UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA MULTIDISCIPLINAR DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL DE CONHECIMENTO

FAUSTO BERNARD MELO SOARES

UTILIZANDO O PROCESSO AUTOMÁTICO DE DESCOBERTA DE CONHECIMENTO PARA CARACTERIZAÇÃO DO PERFIL DAS SUBMISSÕES DOS PESQUISADORES DO INSTITUTO FEDERAL DE SERGIPE: UM ESTUDO DE CASO SOBRE DADOS DO CONGRESSO NORTE NORDESTE DE PESQUISA E INOVAÇÃO

Maceió
2015

FAUSTO BERNARD MELO SOARES

UTILIZANDO O PROCESSO AUTOMÁTICO DE DESCOBERTA DE CONHECIMENTO PARA CARACTERIZAÇÃO DO PERFIL DAS SUBMISSÕES DOS PESQUISADORES DO INSTITUTO FEDERAL DE SERGIPE: UM ESTUDO DE CASO SOBRE DADOS DO CONGRESSO NORTE NORDESTE DE PESQUISA E INOVAÇÃO

Dissertação de mestrado apresentada como requisito parcial para obtenção do grau de mestre em Modelagem Computacional de Conhecimento pela Universidade Federal de Alagoas.

Orientador: Prof. Dr. Aydano Pamponet Machado

Maceió
2015

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

S676u Soares, Fausto Bernard Melo.
Utilizando o processo automático de descoberta de conhecimento para caracterização do perfil das submissões dos pesquisadores do Instituto Federal de Sergipe: um estudo de caso sobre dados do Congresso Norte Nordeste de Pesquisa e Inovação / Fausto Bernard Melo Soares. – 2015.
75 f. : il.

Orientador: Aydano Pamponet Machado.
Dissertação (Mestrado em Modelagem Computacional do Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2015.

Bibliografia: f. 72-74.
Apêndices: f. 75.

1. Pesquisa. 2. Agrupamento de dados. 3. Armazenamento de dados. 4. Ambientes virtuais de aprendizagem. I. Título.

CDU: 004.9:378

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Modelagem Computacional de Conhecimento
Avenida Lourival Melo Mota, Km 14, Bloco 09, Cidade Universitária
CEP 57.072-900 – Maceió – AL – Brasil
Telefone: (082) 3214-1364

MODELAGEM COMPUTACIONAL DE CONHECIMENTO

Membros da Comissão Julgadora da Dissertação de Mestrado de Fausto Bernard Melo Soares, intitulada: "UTILIZANDO O PROCESSO AUTOMÁTICO DE DESCOBERTA DE CONHECIMENTO PARA CARACTERIZAÇÃO DO PERFIL DAS SUBMISSÕES DOS PESQUISADORES DO INSTITUTO FEDERAL DE SERGIPE: UM ESTUDO DE CASO SOBRE DADOS DO CONGRESSO NORTE NORDESTE DE PESQUISA E INOVAÇÃO", apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas, em 4 de setembro de 2015, às 14h00min, no auditório do Instituto de Computação.

COMISSÃO JULGADORA

Prof. Dr. Aydano Pamponet Machado
UFAL – Instituto de Computação
Orientador

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Examinador

Prof. Dr. Silvanito Alves Barbosa
IFS – Instituto Federal de Sergipe
Examinador

Maceió, setembro de 2015.

AGRADECIMENTOS

Agradeço primeiramente a Deus por minha vida e por ter me concedido a oportunidade de viver todo esse desafio e ter chegado a sua conclusão com a sensação de dever cumprido.

A meu pai, Cláudio José Soares, e a minha mãe, Yara Maria Santos Melo, pelo aprendizado, carinho e esforços dedicados a mim ao longo de meus 29 anos. As minhas irmãs Cláudia Danielle e Julielle Luana por fazerem parte dessa caminhada, contribuindo assim para o meu crescimento.

A minha querida esposa Bárbara Prince pelo seu amor, paciência e pelas horas ao meu lado ao longo dessa jornada. Obrigado por não me deixar desanimar e ter contribuído grandemente para que essas palavras finalmente fossem colocadas nessa dissertação. Seu amor me fez forte e foi o ânimo para conclusão dessa etapa em minha vida. Amo você.

Ao Instituto Federal de Sergipe (IFS), pela parceria na promoção desse mestrado e especialmente à Pró-reitoria de Pesquisa e Extensão (PROPEX), pela oportunidade do desenvolvimento desta dissertação através do congresso por ela promovido no ano de 2011.

Ao professor Dr. Aydano Pamponet pela orientação, dedicação e grande ajuda no plano traçado para elaboração desse trabalho, bem como sua execução.

A todos os professores do programa que contribuíram com seu conhecimento e competência ao longo das aulas nos finais de semana aqui no IFS. Conhecimento esse que não serviu apenas para a conclusão do mestrado, mas sim como aprendizado para as circunstâncias da vida.

Ao todos os colegas de turma pelas horas de estudo, finais de semana dedicados aos trabalhos e reuniões para discussões e aprendizado, culminando assim, com a conclusão deste mestrado.

Aos colegas da Diretoria de Tecnologia da Informação (DTI/IFS), pelo apoio em horas difíceis, suporte em momentos de aula e construção da pesquisa que originou esse trabalho.

RESUMO

Em crescente desenvolvimento no Brasil, a atividade de pesquisa vem se tornando cada vez mais presente no cotidiano das instituições de ensino no país. Dessa maneira os Institutos Federais de Educação Ciência e Tecnologia têm adequado sua realidade para atender a essa demanda pela qual passa a sociedade. Isso porque, após a transformação ocorrida no ensino técnico e tecnológico no Brasil, por meio da conversão dos Centros Federais de Educação Tecnológica para Institutos Federais de Educação Ciência e Tecnologia, percebeu-se a necessidade de fortalecer o desenvolvimento da pesquisa nessas instituições, até então voltadas fortemente para a formação de mão de obra. Nesta senda, investimentos têm sido feitos pelo Governo Federal na promoção de programas de incentivo à pesquisa. No Instituto Federal de Sergipe (IFS), em específico, pode-se citar a aplicação de recursos em programas institucionais de importante elevo tais como o Programa Institucional de Bolsas de Iniciação Científica (PIBIC), Programa Institucional de Bolsas de Iniciação em Tecnologia e Inovação (PIBITI) e o Programa Institucional de Apoio a Pesquisa ao Técnico Administrativo (PPTA). Por conta disso, o presente trabalho, utilizando-se do Processo de Descoberta de Conhecimento (KDD) através do algoritmo de agrupamento de dados denominado x-means, propõe, mediante o uso da base de dados de avaliações das submissões de artigos no VI CONNEPI (Congresso Norte Nordeste de Pesquisa e Inovação), ocorrido no ano de 2011, um modelo para o perfil dos trabalhos submetidos nesse congresso pelos pesquisadores do Instituto Federal de Sergipe (IFS) e pelos demais pesquisadores dos Institutos Federais de Educação Ciência e Tecnologia de todo o país. O enfoque precípuo é que, através dos modelos obtidos, se realize uma análise comparativa a fim de mapear características dos pesquisadores dessas instituições e propor, assim, ações em auxílio à tomada de decisão por parte da gestão dessa atividade no IFS, realizada no atual momento pela Pró-Reitoria de Pesquisa e Inovação (PROPEX).

Palavras Chave: Atividade de pesquisa. Processo de descoberta de conhecimento. VI Congresso Norte Nordeste de Pesquisa e Inovação. Agrupamento de dados.

ABSTRACT

In ascending development in Brazil, research activity is becoming more common in everyday life of educational institutions in the country. This way the Federal Institute of Science and Technology Education has adequate its reality to attend this demand through which passes society. Because after the transformation that occurred by technical and technological education in Brazil, through the conversion of Federal Center of Technology Education to Federal Institute of Science and Technology Education, realized the need to strengthen the research development in institution, focused on the formation of labor. So the Federal Government made investments in encouraging research activities programs. In the Federal Institute of Sergipe (IFS) we can mention the application of resources in important institutional programs as the Institutional Program of Initiation Scholarships in Technology and Innovation (PIBIC) and the Institutional Program of Support to Activity Research of Administrative Technician (PPTA). So this dissertation, through of Knowledge-Discovery in Databases (KDD) by means of the clustering algorithm, x-means, proposes by using of the database of evaluations of submitted scientific paper to VI CONNEPI (VI North Northeast Research and Innovation Congress), ocurred in 2011, a model to profile of submitted scientific paper in this congress to researchers of IFS and other researchers of Federal Institute of Science and Technology Education in Brazil. The essential approach is that, through the obtained models, a comparative analysis is performed with the objective to map features of the researchers these institutions and propose actions to the aid to decision-making to IFS, performed by Pró-Reitoria de Pesquisa e Inovação (PROPEX).

Keywords: Research activity. Knowledge-Discovery in Databases VI North Northeast Research and Innovation Congress. Data clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

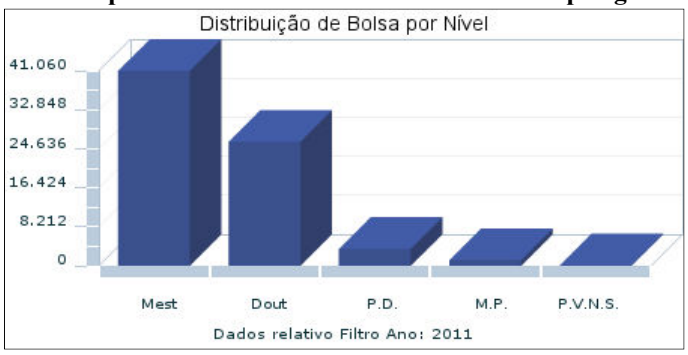
LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

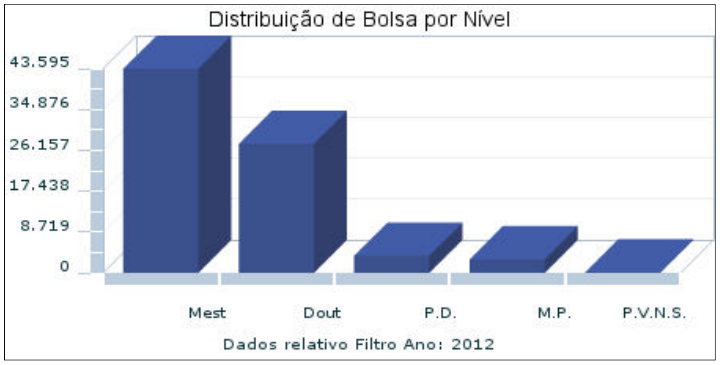
A atividade de pesquisa no Brasil vem crescendo e ocupando um espaço cada vez mais privilegiado nas Instituições de Ensino Superior. Prova disso é que em 2012, foram concedidas pela CAPES (Coordenação de Aperfeiçoamento de Nível Superior), respeitada instituição de investimento à formação de pessoal qualificado no Brasil, aproximadamente 77.900 bolsas nos mais diversos níveis de pós-graduação no país e nas mais diversas áreas de pesquisa. Conforme consta na base de dados da CAPES, o Geocapes[1] (2013), tal número ultrapassa em cerca de 5.000 bolsas o ano anterior, conforme pode ser visto no Gráfico 1 e Gráfico 2.

Gráfico 1- Bolsas fornecidas pela CAPES nos mais diversos níveis de pós-graduação no ano de 2011



Fonte: Geocapes, 2012

Gráfico 2- Bolsas fornecidas pela CAPES nos mais diversos níveis de pós-graduação no ano de 2012



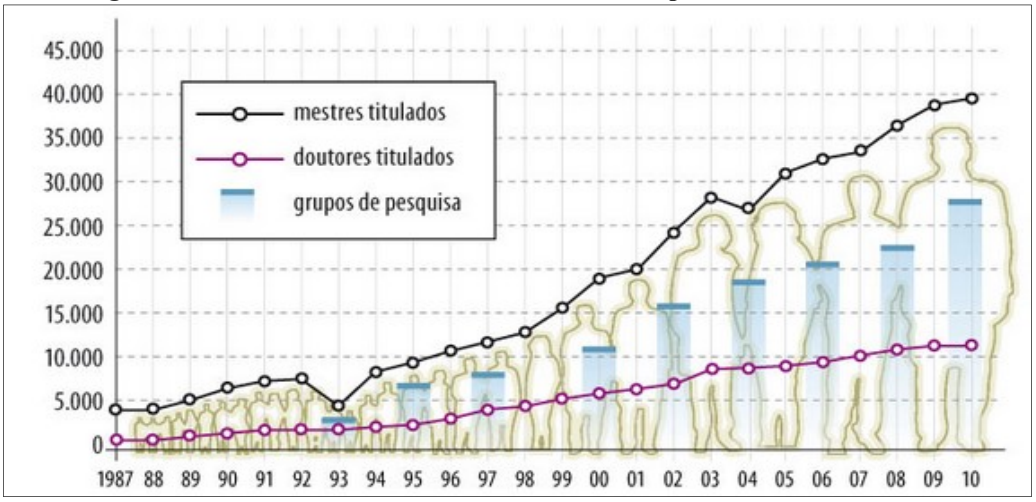
Fonte: Geocapes, 2012

Este número deve-se ao fato do crescimento pelo qual têm passado nos últimos anos as instituições de ensino no país. A procura pela capacitação através do mestrado, doutorado ou pós-doutorado desenvolveu e alavancou fortemente a atividade de pesquisa no Brasil,

[1] Base de dados georreferenciada que disponibiliza informações a respeito da atuação da CAPES no Brasil de acordo com a localização geográfica no país.

conforme visto na Figura 1. Em especial as instituições de ensino público, sejam elas universidades ou institutos federais de educação tecnológica, que em dez anos (2002 a 2012) tiveram crescimento[2] considerável em número de unidades e Campi e, por consectário, começaram a despontar fortemente no tocante à produção de pesquisa, despontando inclusive no cenário internacional, através, por exemplo, de publicações em periódicos de alto conceito no panorama mundial (BRASIL *et al*, 2012).

Figura 1 - Número de doutores e mestres formados por ano de 1989 até 2010



Fonte: Ministério da Educação e Cultura, 2012

Ao contrário do que se pensa, nos dias de hoje a atividade de pesquisa não é exercida somente por uma minoria em universidades ou grandes centros em nosso país. Tal atividade também está sendo desenvolvida em diversos laboratórios e salas de aula das universidades ou institutos federais de educação tecnológica de pequenos centros, inclusive do próprio estado de Sergipe, conforme visto no anuário estatístico[3] da Universidade Federal de Sergipe e no Relatório de Autoavaliação Institucional[4] do IFS em 2012. Este fato pode ser comprovado em dados do crescimento de investimento do governo, por meio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através de bolsas de pesquisa nos últimos anos nas academias brasileiras, conforme Gráfico 3.

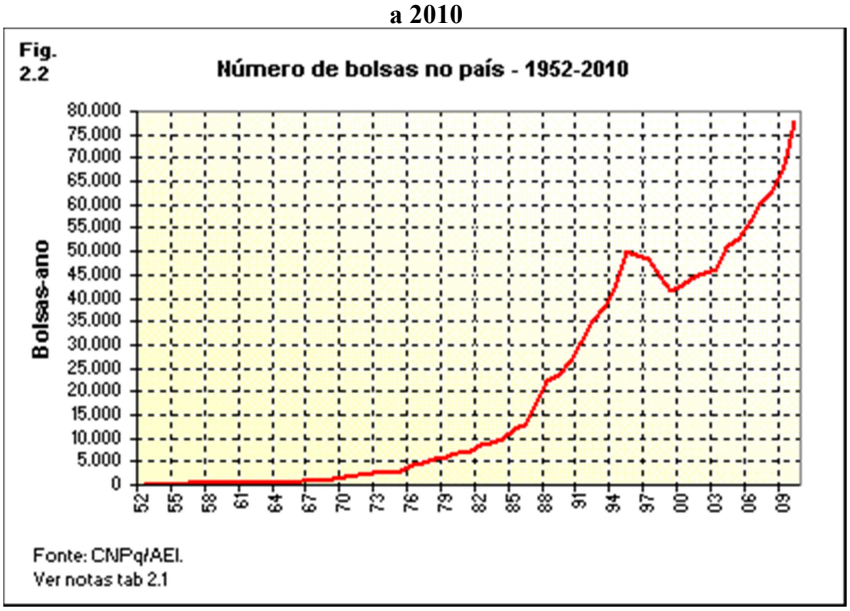
---

2 Em 2002 havia 43 universidades e 148 campi, oito anos depois o Brasil possuía 59 instituições públicas de ensino superior e 247 campi. Existe a previsão do MCTI para 2014, que o Brasil tenha 63 universidades pelos 321 campi universitários e 362 institutos de educação profissional e tecnológica (BRASIL et al, 2012).

3 O anuário estatístico aqui citado refere-se aos anos de 2009 a 2011.

4 Tal relatório é desenvolvido por uma comissão própria de avaliação (CPA). Este pode ser conferido na íntegra em: http://www.ifs.edu.br/index.php?option=com_content&view=article&id=1186&Itemid=68.

Gráfico 3- Quantitativo de bolsas de pesquisa concedidas pelo governo federal por meio do CNPq de 1952 a 2010



Fonte: CNPq, 2011

Assim, atrativo ao pesquisador, o incentivo por parte do governo na modalidade de bolsa vem se tornando um grande aliado ao desenvolvimento da pesquisa no país, alavancando, dessa forma, a produção de conhecimento nas mais diversas áreas de pesquisa, seja para arcar com os custos da pesquisa desenvolvida, seja através da compra de bibliografia especializada sobre o tema, ou ainda com gastos de material e equipamento para o seu progresso.

Dessa forma, o governo através de diversas agências de fomento vem incentivando esta prática, tornando a pesquisa uma realidade nos dias de hoje nas instituições de ensino de todo o Brasil. Dados do CNPq[5] revelam que, no ano de 2012, o próprio Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) investiu cerca de R$ 432.512.725,00 para auxiliar no desenvolvimento de projetos de pesquisa no país. Além disso, as próprias instituições de ensino destinam recursos próprios para incentivar e fomentar a pesquisa junto a sua comunidade acadêmica.

Nesse contexto, somente o Instituto Federal de Sergipe, segundo dados do Relatório de Gestão do Exercício 2012 da Pró-reitoria de Pesquisa e Extensão, empregou cerca de R$ 726.306,80 em fomento à atividade de pesquisa no ano de 2012, procurando, assim, ampliá-la e desenvolvê-la na própria Instituição, proporcionando dessa maneira, ao professor e, mais recentemente, ao técnico-administrativo, a oportunidade de envolvimento nesta atividade nos mais diversos Campi pelo Estado de Sergipe.

---

5 Conforme consta em http://fomentonacional.cnpq.br/dmfomento/home/fmtvisualizador.jsp.

1.1 Motivação e Definição do Problema

Mediante o exposto, notoriamente percebe-se que vêm sendo empregados pelo governo federal, agências financiadoras, empresas de economia mista, esforços na melhoria da prática da atividade de pesquisa no país. Desta forma, as instituições que recebem e aplicam tal incentivo financeiro devem balizar a melhor forma para aplicação dessa verba ao desenvolvimento e fomento da atividade de pesquisa.

Também se sabe que as instituições de ensino possuem suas dificuldades em traçar o melhor caminho para investimento do recurso destinado a pesquisa, o que por vezes é majorado pela falta de um diagnóstico mais próximo da realidade no que se diz respeito a esta atividade na instituição.

Assim, a falta de informação mais significativa acerca da produção de seus pesquisadores, bem como informações que produzam um diagnóstico mais preciso faz com que as instituições de ensino no Brasil sofram com uma frágil gestão no que diz respeito à pesquisa. O conhecer do andamento da atividade de pesquisa na instituição possibilita a definição de estratégias que viabilizem o seu desenvolvimento, potencializando assim a sua prática.

Não diferentemente do que fora citado, o Instituto Federal de Sergipe (IFS), tomado como estudo de caso no presente trabalho, vem enfrentando um desafio no decorrer dos últimos anos, mais precisamente no último triênio, de proporcionar o desenvolvimento da pesquisa, atividade cuja qual não se possui grandes informações ou diagnósticos na instituição, em vista das diversas mudanças propostas à rede de ensino federal e tecnológica, dentre elas aumentar notoriamente a produção de atividades de pesquisa e extensão.

1.2 Justificativa

Durante os últimos anos, a mudança de concepção acadêmica das escolas técnicas tem exigido investimento no campo da pesquisa e extensão dessas instituições, as quais já se caracterizam pela excelência no ensino técnico e tecnológico no país. O desafio de impulsionar a atividade de pesquisa e extensão ganhou elevo nacional na medida em que o Governo Federal propôs uma mudança conceptiva a toda rede federal de ensino técnico e tecnológico: da educação técnica tradicionalista evolui-se para o orbe do desenvolvimento cognitivo contemporâneo por intermédio da pesquisa e extensão.

O Instituto Federal de Sergipe, instituição de estudo de caso deste trabalho, vem a cada ano demandando esforços para o crescimento adequado da sua atividade de pesquisa e extensão através de diversas ações de gestão e acadêmicas. A exemplo, pode-se citar a implementação de laboratórios didáticos, lançamento de editais anuais de incentivo ao desenvolvimento de projetos de pesquisa na instituição junto à oferta de dezenas de bolsas a alunos, professores e técnicos administrativos, oferta anual de cursos de extensão para comunidade interna e externa, não se esquecendo da promoção cada vez mais intensa da SNCT – Semana de Ciência e Tecnologia, com apresentação de frutos dos projetos desenvolvidos na instituição, além da promoção do VI CONNEPI (Congresso Norte Nordeste de Pesquisa e Inovação) no ano de 2011, o que gerou grande entusiasmo em pesquisadores (seja docentes ou técnicos administrativos), contribuindo de forma positiva para potencializar a produção acadêmica e o desenvolvimento desta atividade na instituição.

Ocorre que diversas são as dificuldades enfrentadas por uma instituição essencialmente criada nos moldes tradicionais de ensino. Tais dificuldades enfrentadas fazem com que o foco deste trabalho seja auxiliar a Pró-reitoria de Pesquisa e Extensão na sua tarefa de proporcionar a atividade de pesquisa a toda a comunidade que faz parte do Instituto Federal de Sergipe. E tal auxílio se manifesta em conhecer o perfil das submissões realizadas no VI CONNEPI pelos pesquisadores que a instituição abriga, suas habilidades e fragilidades, questões estas que serão discutidas e subsidiadas através de um modelo computacional que se utiliza da mineração de dados como ferramenta de descoberta de conhecimento e auxílio na tomada de decisões à gestão da pesquisa no Instituto Federal de Sergipe.

Através do processo de descoberta de conhecimento busca-se por características nos dados das avaliações das submissões de pesquisadores do país, por meio do uso de um algoritmo de agrupamento de dados. Mediante os grupos gerados nesse processo é possível obter informações que descrevam de maneira mais aprofundada os perfis resultantes da mineração, visto que os dados são agrupados em clusters de acordo com métricas de similaridade que estabelecem, dessa forma um relacionamento entre os dados e o grupo ao qual pertence.

Nesta senda, a utilização apenas da estatística descritiva nesse processo, seja por meio de medidas de tendência central ou medidas de dispersão, não elucida o conhecimento oculto existente na base de dados. Fez-se, por isso, a opção pelo uso de uma técnica de aprendizagem não supervisionada, cabendo a esta o estabelecimento de padrões e características dos dados selecionados nesse processo por meio da criação de grupos que abriguem e caracterizem os dados das submissões dos pesquisadores no congresso em estudo.

Indiscutivelmente, elucidar características dos pesquisadores do IFS, detectando suas potencialidades e fragilidades, auxiliará sobremaneira a PROPEX em sua tarefa de planejamento e gestão da pesquisa no Instituto Federal de Sergipe, fortalecendo e oferecendo, desta maneira, uma constante evolução da atividade de pesquisa e extensão no Instituto.

1.3 Objetivos

Apoiando-se na ideia de se auxiliar a gestão e o planejamento direcionado à pesquisa, de modo a direcionar esforços que a façam evoluir e se consolidar cada vez mais no cenário nacional e, procurando dar apoio às decisões tomadas pela PROPEX, por meio da descoberta de conhecimento, pode-se destacar dentre os objetivos do presente trabalho, os seguintes:

- Mapear as características dos trabalhos enviados ao congresso por pesquisadores do IFS, detectando, assim, suas potencialidades e fragilidades, sendo possível identificar por meio do processo de descoberta de conhecimento, através do uso da atividade de agrupamento de dados, padrões na produção científica dos pesquisadores do Instituto Federal de Sergipe, realizando uma análise comparativa com pesquisadores da rede federal de ensino técnico e tecnológico de todo o país;
- Apoiar a PROPEX em seu planejamento de pesquisa, fornecendo informações acerca da produção dos pesquisadores do IFS no congresso para que assim ações de melhoria no desenvolvimento da atividade de pesquisa na Instituição sejam tomadas pela pró-reitoria. Conhecendo as características dos pesquisadores do Instituto Federal de Sergipe, a PROPEX encontra balizas técnicas para a tomada de decisões que influenciam no desenvolvimento da atividade pesquisa na instituição. Nesse contexto, é possível fornecer meios para definição e melhor oferta de seus recursos e esforços empenhados em pesquisa;
- Promover um estudo comparativo entre a produção dos pesquisadores do IFS e os demais pesquisadores do país baseado na submissão de artigos no VI CONNEPI.

Buscando alcançar tais objetivos, utilizou-se como subsídio a análise dos artigos submetidos no VI CONNEPI, evento ocorrido no ano de 2011, no qual concentrou a atenção de pesquisadores do todo o país, seja com submissões de trabalhos, seja com a atividade de avaliação dos artigos propostos. Para o último caso, o pesquisador deveria possuir no mínimo o grau de mestre para a área cuja qual estava pleiteando realizar avaliações.

A escolha pelo VI CONNEPI, o terceiro Congresso Norte Nordeste de Pesquisa e Inovação ocorrido após a transformação dos até então CEFET's (Centros Federais de

Educação Tecnológica) em Institutos Federais, advém da importância deste evento para toda a rede de educação tecnológica, sobretudo no aspecto da submissão de pesquisas desenvolvidas ao longo do ano de 2011.

Com elevo, é possível observar a adaptação do agora IFS a sua nova realidade como Instituto Federal de Educação Tecnológica, visto o interstício de pouco mais de 2 anos desde sua criação e o evento em 2011. Assim, com o intuito de avaliar essa mudança ocorrida nos antigos centros federais, o presente trabalho traz uma análise do que está sendo desenvolvido pelo corpo de pesquisadores dos institutos federais do país, seja ele composto por docentes ou técnicos administrativos.

1.4 O Desenvolvimento da Proposta

O presente projeto foi assistido pela utilização do Processo de Descoberta de Conhecimento (*Knowledge Discovery in Databases – KDD*), uma metodologia que auxilia na busca por conhecimento oculto em base dados, com emprego da mineração de dados, utilizando-se algoritmos de aprendizagem de máquina, tanto supervisionada quanto não-supervisionada (KUMAR *et al*, 2009).

Neste processo de descoberta de conhecimento, a obtenção e o tratamento da base de dados a ser utilizada é essencial como ponto de partida. Por primeiro, faz-se necessária a extração da informação contida no banco de dados do sistema que cuidou do gerenciamento das submissões dos trabalhos. Após isso, é estabelecido o uso de algoritmos de agrupamento de dados para tornar possível a obtenção de grupos dentro do conjunto de dados obtidos na submissão dos artigos no evento e, assim, poder estabelecer padrões nesses dados; em outras palavras, procurar-se-á definir um relacionamento entre os pesquisadores que submeteram seus trabalhos ao evento por meio de suas características de produção científica, características essas que foram definidas pela comissão científica do Congresso em 2011 e, assim, utilizadas na realização da avaliação dos artigos submetidos ao evento.

De posse dos grupos formados pelo algoritmo, é possível uma leitura que elucide muito mais do que a estatística descritiva (KUMAR *et al*, 2009). É possível investigar a fundo a relação entre os parâmetros de produção científica, criados para o VI CONNEPI, e retirar informações de grande valia para o Instituto Federal de Sergipe. Informações que ultrapassem um diagnóstico superficial e inicial dos pesquisadores, de modo a ler nas entrelinhas da base de dados de avaliação dos artigos submetidos, encontrando padrões de potencialidades e fragilidades na produção científica do evento. Esta é uma das principais funções do Processo

de Descoberta de Conhecimento aqui aplicado, que assim subsidiará ferramentas de apoio a tomada de decisão por parte da gestão para melhor aplicação e desenvolvimento da atividade de pesquisa.

1.5 Estrutura do Trabalho

Organizado em 5 capítulos, o presente trabalho em seu Capítulo 2 trata da questão que envolve toda a problemática e situa no contexto nacional o desenvolvimento da pesquisa no Instituto Federal de Sergipe. Neste capítulo também é referenciado aspectos históricos da Pró-Reitoria de Pesquisa e Extensão (PROPEX) no Instituto Federal de Sergipe, bem como são elucidadas pontos referentes às ações por ela promovidas, tais como: PIBIC (Programa Institucional de Bolsas de Iniciação Científica), PIBICJR (Programa Institucional de Bolsas de Iniciação Científica Júnior), PIBITI (Programa Institucional de Bolsas de Iniciação Cientifica em Desenvolvimento Tecnológico e Inovação), PPTA (Programa Institucional de Apoio à Pesquisa ao Técnico-administrativo da Educação), não se esquecendo do esforço desempenhado em seguir uma orientação nacional na produção da SNCT (Semana de Ciência e Tecnologia).

Ainda no Capítulo 2, é expressa a teoria útil ao trabalho que envolve o Processo de Descoberta de Conhecimento e a tarefa de análise de agrupamento de dados, ponto fundamental de auxílio ao progresso com êxito da pesquisa em tela. Nesse ponto, discute-se o Processo de Descoberta de Conhecimento, uma metodologia eficiente na busca e extração do conhecimento implícito na base de dados examinada, suas fases e como representa auxílio indispensável a este trabalho, bem como os algoritmos k-means e x-means aqui utilizados para a criação dos agrupamentos, possibilitando a análise e investigação dos mesmos.

A metodologia, trazida no Capítulo 3, relata os passos percorridos neste trabalho para propor o modelo computacional que possibilite uma análise consistente das informações obtidas na submissão e avaliação dos artigos apresentados ao VI CONNEPI. Aqui se relata o processo desde a obtenção dos dados por meio do sistema de submissões desenvolvido no Instituto Federal de Sergipe no ano de 2011, seguindo pelo tratamento de toda essa informação bruta acumulada, a fim de ter uma base limpa e que facilite a aprendizagem de máquina aplicada, passando pelo processo de mineração de dados no momento da construção dos grupos através do uso do algoritmo de *clustering* x-means, e finaliza-se gerando assim subsídio para discussão dos resultados no capítulo seguinte.

No Capítulo 4, intitulado *Resultados e análises dos agrupamentos*, são exibidos os

resultados do que foi proposto no capítulo anterior. Aqui são concretizados todos os passos relatados no Capítulo 3 exibindo assim uma descrição das bases de dados trabalhadas por meio da estatística descritiva, que traz um diagnóstico da base de dados após a sua obtenção e limpeza. Também são trazidos, neste capítulo, os resultados da etapa de processamento do KDD. Quanto à mineração de dados em si, pode-se ver dois momentos. Inicialmente através da divisão em grupos utilizando-se do algoritmo explorado, bem como os centroides de cada grupo criado após a divisão dos dados em grupos rotulados pelo processo de aprendizagem não-supervisionada. No segundo momento, um grupo em específico será subdividido criando assim subsídio para explorar as características e comportamentos de cada agrupamento gerado no momento anterior.

Finalmente, no Capítulo 5, são trazidas as conclusões a cerca da solução proposta. Aqui são traçados os perfis dos trabalhos submetidos ao VI CONNEPI pelos pesquisadores do Instituto Federal de Sergipe e pelos demais Institutos Federais de Educação, Ciência e Tecnologia que participaram do congresso, obtidos através do processo de descoberta de conhecimento. Também é realizada uma comparação entre os dois perfis buscando elucidar possíveis fragilidades e potencialidades na atividade de pesquisa no IFS. Além disso, o presente capítulo, através do auxílio da aprendizagem de máquina utilizando-se as ações da pró-reitoria no processo de tomada de decisão, propõe indicar direcionamentos no processo de gestão da pesquisa na instituição. E, por fim, são fomentados, neste capítulo, trabalhos futuros que permeiem o que foi trabalhado nesta pesquisa, além das devidas conclusões a respeito do que foi inicialmente proposto.

2 CONTEXTUALIZAÇÃO E FUNDAMENTAÇÃO TEÓRICA

O presente capítulo traz nas linhas que se seguem a situação em que se encontra o desenvolvimento da atividade de pesquisa no Instituto Federal de Sergipe – IFS, mediante um histórico trazido desde o ano de 2009. Tal levantamento contém dados e informações sobre a política de desenvolvimento de pesquisa na Instituição, bem como investimentos nesta atividade e retorno trazido pela comunidade de seus pesquisadores: alunos, docentes e técnicos administrativos, para seu próprio fortalecimento acadêmico.

Traz também um levantamento sobre o processo de descoberta de conhecimento em bases de dados. Objetiva-se mostrar do que se trata esse processo tão poderoso nos dias atuais e como trabalham os algoritmos de aprendizagem não-supervisionada na obtenção de conhecimento em bases de dados onde visualmente essa tarefa não é tão simples.

Assim, será trazido aqui o que é esse processo de descoberta de conhecimento, conhecido também como KDD (Knowledge Discovery in Databases), bem como a atividade de mineração de dados e os algoritmos de agrupamento de dados conhecidos como k-means e x-means.

2.1 O Cenário da Pesquisa no Instituto Federal de Sergipe

Recentemente convertido à Instituto Federal de Educação Tecnológica[6], o IFS vem galgando passos rumo ao seu desenvolvimento no estado de Sergipe e também no cenário nacional. Tais passos vêm sendo apoiados através de ações do Governo Federal para manutenção e evolução de sua atuação no país. Dentre as ações do governo, pode-se citar a concessão de bolsas de fomento às atividades de pesquisa e extensão geridas no Instituto Federal de Sergipe pela Pró-Reitoria de Pesquisa e Extensão (RGE2012 – PROPEX, 2013)[7].

Segundo o Regulamento de Pesquisa e Extensão do Instituto Federal de Sergipe (2011), cabe ainda a esta Pró-reitoria o trabalho de melhor investir esse recurso disponibilizado pelo governo ao IFS em seus mais diversos programas de incentivo à pesquisa na instituição. Tais programas são: PIBIC (Programa Institucional de Bolsas de Iniciação Científica), PIBICJr (Programa Institucional de Bolsas de Iniciação Científica Júnior), PPTA

---

[6] A conversão de Centro Federal de Educação Tecnológica de Sergipe para Instituto Federal de Sergipe ocorreu no ano de 2008 através da Lei n.°11.892 de 29 de dezembro de 2008.

[7] Relatório de Gestão do Exercício 2012 da Pró-reitoria de Pesquisa e Extensão (PROPEX) do IFS.

(Programa Institucional de Apoio à Pesquisa ao Técnico Administrativo) e Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI).

Conforme consta no Relatório de Gestão do Exercício 2012 da PROPEX, além dos programas supracitados desenvolvidos pela instituição, cabe à PROPEX acompanhar e participar dos editais disponibilizados pelas agências de fomento a pesquisa e assim conseguir aprovar os projetos provenientes dos pesquisadores do IFS. Dentre as agências de fomento cujas quais o IFS possui projetos financiados pode-se citar a Fundação de Apoio à Pesquisa do Estado de Sergipe (FAPITEC), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Nível Superior (CAPES), através dos programas Jovens Talentos/CAPES, PIBIC/CNPq, PIBITI/CNPq, PIBICJr./FAPITEC, PIBIC/FAPITEC e PIBITI/FAPITEC.

2.1.1 Programas de pesquisa desenvolvidos no Instituto Federal de Sergipe

Os programas de pesquisa são ofertados pelo IFS, através da PROPEX, ao mais diverso público da instituição, composto de alunos de ensino técnico subsequente ou integrado, PROEJA, alunos de ensino superior, dos docentes da instituição das mais diversas áreas de atuação e ainda os técnicos administrativos, através de um programa especial dedicado exclusivamente a essa categoria (RPE – PROPEX, 2011).

Para participação nos programas ofertados pela instituição, o projeto a ser executado deve ser submetido à Coordenação de Apoio à Pesquisa (CAP) pelo coordenador e responsável pelo mesmo, em período predeterminado pela PROPEX. Tal coordenador deverá ser um docente ou técnico administrativo e possuir preferencialmente, mas não exclusivamente, o título de doutor, podendo o mestre também submeter e ter sua proposta aprovada (RPE – PROPEX, 2011).

2.1.1.1 Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Focado na iniciação científica, o presente programa, implementado através da concessão de bolsas nos mais diversos Campi do IFS, procura atingir alunos de nível técnico subsequente e de nível superior, além dos professores da instituição na função de orientador do projeto.

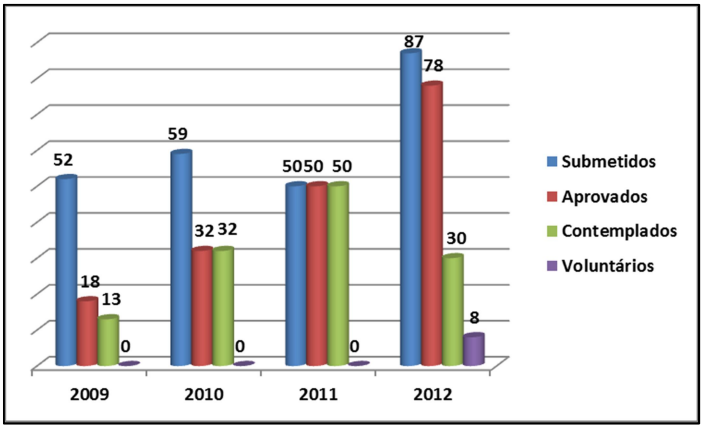
Dentre os objetivos desse programa encontra-se o de contribuir para a formação e iniciação científica do aluno, tornar institucional a ação da pesquisa no âmbito dos Campi do

IFS, fortalecer os grupos de pesquisa presentes na instituição, além de gerar conhecimento e buscar soluções de acordo com a necessidade dos diversos segmentos produtivos do estado de Sergipe (RGE2012 – PROPEX, 2013).

No ano de 2012 houve um investimento nesse programa de R$ 210.000,00 (duzentos e dez mil reais), proveniente de verba própria do Instituto Federal de Sergipe, entre bolsas para alunos, bolsas para coordenador de projeto, auxílio financeiro para custeio de despesas com projeto no que diz respeito a material de consumo, serviços de terceiros, deslocamento, aquisição de livros e periódicos. Tal valor é superior ao empregado pelo IFS no ano de 2011, que teve como despesa neste programa no referido ano o valor de R$ 35.000,00 (trinta e cinco mil reais), entre despesas com bolsas para discentes e docentes (RAI – CPA, 2012).

Assim, é possível perceber o esforço de investimento financeiro direcionado à atuação deste programa na instituição, percebendo-se um aumento da oferta e execução de projetos PIBIC nos últimos anos, aumento este de 230% no quantitativo de projetos aprovados contemplados com recurso, além de 67,3% da quantidade de projetos submetidos no período, conforme Gráfico 4.

Gráfico 4 - Quantidade de projetos submetidos, aprovados, contemplados e voluntários no PIBIC do IFS nos anos de 2009 a 2012



Fonte: RGE2012 – PROPEX, 2013

2.1.1.2 Programa Institucional de Bolsas de Iniciação em Tecnologia e Inovação (PIBITI)

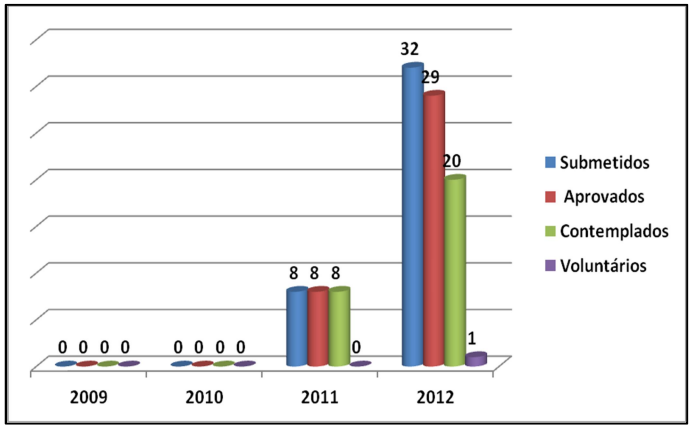
Presente no Instituto Federal de Sergipe e de responsabilidade da Coordenadoria de Ciência e Tecnologia, o PIBITI tem como função estabelecer e estimular os docentes e alunos de nível técnico subsequente e nível superior do Instituto à prática de atividades de pesquisa no tocante ao desenvolvimento tecnológico e inovação. Faz-se instigar o pensamento do pesquisador envolvido nesse programa de forma empreendedora, focado no desenvolvimento

criativo e inovador da pesquisa. Desta maneira, estimula-se uma maior aproximação das empresas em suas soluções buscando desenvolver um vínculo mais forte entre o IFS e a comunidade (RGE2012 – PROPEX, 2013).

O investimento no ano de 2012 com este programa atingiu a marca de R$ 150.000,00 (cento e cinquenta mil reais), cujo qual contemplou por oito meses o custeio de bolsas para alunos e também docentes envolvidos nos 20 projetos de pesquisa e inovação aprovados e contemplados com bolsas pela Pró-Reitoria, através da Coordenação de Apoio à Pesquisa. Além disso, o programa possuiu, assim como o PIBIC, uma verba de custeio para despesas do projeto, conforme item 2.1.1.1 (RGE2012 – PROPEX, 2013).

Em comparação ao ano anterior, 2011, cujo valor destinado ao programa foi de R$ 5.600,00 (cinco mil e seiscentos reais), percebe-se um aumento considerável de investimentos específico a essa ação, que no IFS foi implantada no ano de 2011 (RAI – CPA, 2012). De igual modo, pode-se perceber notável crescimento na quantidade de trabalhos que foram submetidos, aprovados e contemplados em relação ao ano de implantação, conforme Gráfico 5.

Gráfico 5- Quantidade de projetos submetidos, aprovados, contemplados e voluntários no PIBITI do IFS nos anos de 2009 a 2012



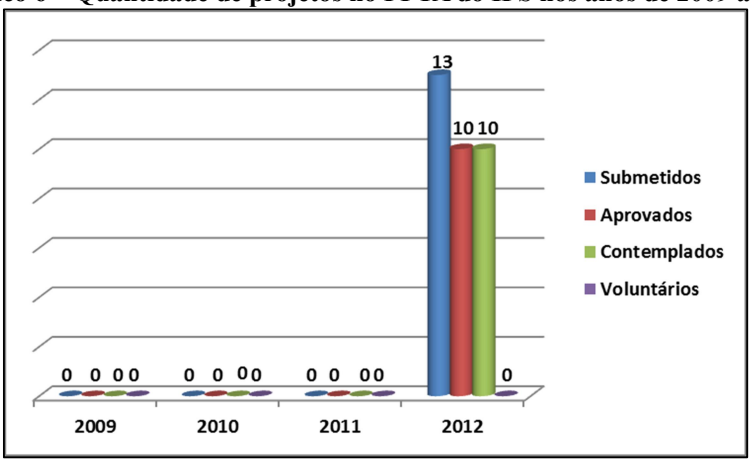
Fonte: RGE2012 – PROPEX, 2013

2.1.1.3 Programa Institucional de Apoio à Pesquisa ao Técnico Administrativo (PPTA)

Novidade no ano de 2012 no Instituto Federal de Sergipe, o Programa Institucional de Apoio à Pesquisa ao Técnico Administrativo (PPTA), recebeu um investimento de R$ 29.500,00 (vinte e nove mil e quinhentos reais) para o desenvolvimento de projetos de pesquisa por técnicos administrativos da instituição, através de bolsas para o custeio com despesas do projeto, conforme dados representados no Gráfico 6 (RGE2012 – PROPEX,

2013). O objetivo precípuo desta ação é o incentivo à participação do corpo de técnicos administrativos da instituição nos mais diversos grupos de pesquisas presentes no IFS e de responsabilidade da PROPEX (RPE – PROPEX, 2011).

Gráfico 6 – Quantidade de projetos no PPTA do IFS nos anos de 2009 a 2012



Fonte: RGE2012 – PROPEX, 2013

2.1.1.4 Programa Institucional de Bolsas de Iniciação Científica Júnior (PIBIC JR)

Semelhante ao PIBIC[8], o Programa Institucional de Bolsas de Iniciação Científica Júnior (PIBIC JR), criado em junho de 2012, tem o objetivo de fomentar a pesquisa na instituição no tocante à iniciação científica, servindo como um passo importante a alunos e professores no desenvolvimento inicial de suas atividades de pesquisa na instituição. Diferente do PIBIC, o PIBIC JR concentra esforços em um público bem definido: alunos do ensino técnico integrado[9], do ensino técnico subsequente e do PROEJA[10], juntamente aos professores da instituição (RGE2012 – PROPEX, 2013).

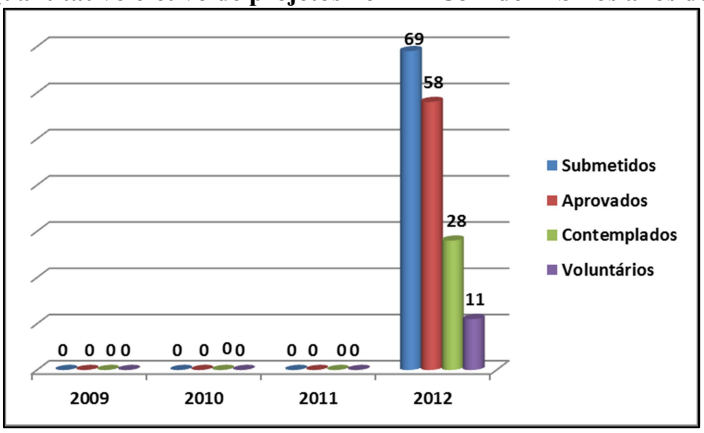
No ano de sua criação, em 2012, foram ofertadas à comunidade 30 bolsas aos alunos no período de 8 meses de execução do projeto, 30 bolsas por igual período ao docente coordenador de cada projeto, além do recurso de apoio financeiro para cada projeto executado. O montante então direcionado a esse programa somou a quantia de R$ 137.800,00 (cento e trinta e sete mil e oitocentos reais), sendo não utilizado, dessa quantia, o valor de R$ 45.200,00 (quarenta e cinco mil e duzentos reais), por conta de não continuidade de alguns projetos ou impossibilidade de recebimento da bolsa visto que no IFS, conforme o

---

8 Conforme descrição do item 2.1.1.1 do presente capítulo.
9 Realização simultaneamente do curso técnico integrado e do ensino médio.
10 Programa de Educação de Jovens e Adultos.

Regulamento de Pesquisa e Extensão proposto em 2011, cada pesquisador não pode receber mais de uma bolsa de pesquisa, mesmo que execute dois ou mais projetos. Assim, totalizou-se o quantitativo efetivo disponível no Gráfico 7 para este programa (RGE2012 – PROPEX, 2013).

Gráfico 7- Quantitativo efetivo de projetos no PIBICJR do IFS nos anos de 2009 a 2012



Fonte: RGE2012 – PROPEX, 2013

Por todo exposto, percebe-se um investimento crescente direcionado à atividade de pesquisa do IFS, além de novos programas e ideias que estão sendo desenvolvidos procurando melhorar a qualidade das ações de pesquisa, seja esta desenvolvida por docentes, alunos ou técnicos administrativos da instituição.

2.1.2 O VI Congresso Norte Nordeste de Pesquisa e Inovação - VI CONNEPI

Promovido pela Rede Norte Nordeste de Educação Profissional e Tecnológica e pela Secretaria de Educação Tecnológica (SETEC) do Ministério da Educação, o Congresso Norte e Nordeste de Pesquisa e Inovação (CONEPPI) tem sido visto como um evento de grande importância para a rede de educação tecnológica.

No ano de 2011, realizado através de uma parceria do Instituto Federal do Rio Grande do Norte e o Instituto Federal de Sergipe, o VI Congresso Norte e Nordeste de Pesquisa e Invocação chamou especial atenção de todos os pesquisadores do IFS e Institutos Federais de Educação Tecnológica do país. Com o tema "Tecnologia inovadora sustentável: Ações afirmativas, equidade e promoção da qualidade de vida" (http://connepi2011.ifs.edu.br/, 2011), o evento propôs constituir discussões e a troca de conhecimento científico e tecnológico entre os pesquisadores do país nas mais diversas áreas de conhecimento, denominadas de áreas temáticas.

Conforme consta no *site* oficial do evento (http://connepi2011.ifs.edu.br/), dentre as mais diversas áreas temáticas do VI CONNEPI, juntamente com as suas subáreas, pode-se citar:

i. Ciências Agrárias: Agronomia, Ciências e Tecnologia de Alimentos, Engenharia de Pesca, Medicina Veterinária e Zootecnia, Recursos Florestais e Engenharia Florestal.

ii. Ciências Exatas da Terra: Ciências da Computação, Física, Geociências, Matemática, Probabilidade e Estatística, Química.

iii. Engenharia: Engenharia Civil, Engenharia Sanitária, Engenharia Química, Engenharia Elétrica, Engenharia Mecânica, Engenharia de Produção.

Dessa maneira, diversos pesquisadores do país puderam propor seus trabalhos na área que mais se adequasse a sua pesquisa, por meio de um artigo científico, devidamente avaliado por mestres e doutores do todo o país, de modo a consolidar a difusão de conhecimento e ideias propostas pelo evento, divulgando, assim, a educação tecnológica junto à sociedade (http://connepi2011.ifs.edu.br).

2.1.3 Semana Nacional de Ciência e Tecnologia (SNCT)

Aderindo à Semana Nacional de Ciência e Tecnologia, o Instituto Federal de Sergipe realiza, desde o ano de 2011, atividades voltadas à difusão e à apropriação social de conhecimentos científicos e tecnológicos (RGE2012 – PROPEX, 2013), atividades estas voltadas a uma temática central proposta pelo governo federal anualmente, estimulando, assim, o desenvolvimento de ações de pesquisa nas instituições de ensino em todo o Brasil, simultaneamente, em determinada época do ano, desde o ano de 2004 (semanact.mct.gov.br/index. php/content/view/5391.html).

No IFS, palestras, apresentação de trabalhos e fóruns de discussão têm sido formados e expostos nos anos de sua participação no evento. Tal medida vem desenvolvendo a pesquisa no meio acadêmico da instituição, que por anos teve como foco principal tão somente a preparação de jovens para o mercado de trabalho a fim de atender uma necessidade imediata do estado de Sergipe, sem tanto envolvimento com a pesquisa e extensão (RAI – CPA, 2012).

2.2 O Processo de Descoberta de Conhecimento e os Algoritmos de Agrupamento de Dados

2.2.1 O processo de descoberta de conhecimento

Nos dias atuais, o avanço do mundo moderno tem mostrado que o homem deve estar cada vez mais preparado para agir sobre os desafios que lhe são propostos. Assim, os dados produzidos e armazenados pelas empresas atualmente possuem valiosas informações que vêm recebendo especial atenção. O conjunto dessas informações, se convertida em conhecimento, torna-se importante ferramenta de balizamento nas mais diversas áreas em que o homem atua (HAN; KAMBER, 2006). Seja na análise de mercado, detecção de fraudes em cartões de crédito ou transações, seja na área da medicina quando da previsão de doenças desenvolvidas por determinado agente, ou nas mais diversas áreas, onde o conhecimento adquirido seja útil no processo de descoberta de conhecimento (KUMAR *et al*, 2009).

O processo para obtenção de conhecimento através de dados armazenados envolve fases que devem ser cumpridas, que vão desde a coleta e armazenamento de dados, passando pela aplicação de técnicas de aprendizagem de máquina, utilizando-se da inteligência artificial, até a análise dos resultados obtidos para extração do conhecimento desejado e que motivou a sua aplicação (KUMAR *et al*, 2009).

Assim, segundo Fayyad *et al* (1996), pode-se definir como KDD o processo pelo qual, através da análise em conjunto de dados, busca-se padrões e tendências, sendo a mineração de dados sua principal etapa no processo, visto que por meio de algoritmos desenvolvidos especialmente para esse fim obtém-se conhecimento e informações implícitas nos dados submetidos.

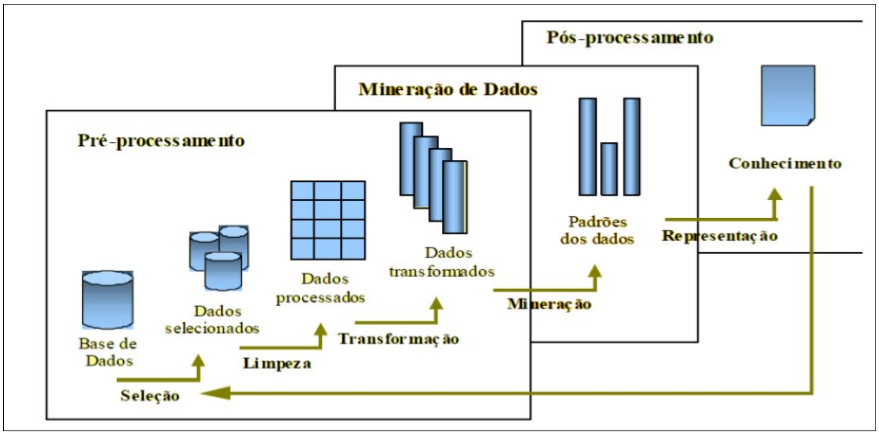
O Processo de Descoberta de Conhecimento é estabelecido sobre algumas etapas a fim de que o conhecimento seja extraído satisfatoriamente dos dados que estão armazenados (HAN; KAMBER, 2006). Segundo Kumar *et al* (2009), as etapas estabelecidas nesse processo, conforme Figura 2, podem ser divididas em:

i. Pré-processamento: etapa inicial que consiste na seleção de recursos, redução de dimensionalidade, normalização e criação de subconjunto de dados. O objetivo desse passo é transformar os dados de entrada para os processos subsequentes. Tal transformação envolve limpeza dos dados, fusão de dados de diversas fontes, observações duplicadas e escolha de registros com informações relevantes ao processo (KUMAR *et al*, 2009);

ii. Processamento: fase de aplicação efetiva da atividade de mineração de dados. É nesse momento que algoritmos de aprendizagem de máquina são submetidos ao processo em busca do conhecimento implícito nos dados, através de tarefas de classificação, agrupamento ou sumarização (FAYYAD, 1996);

iii. Pós-Processamento: Etapa final do processo de descoberta de conhecimento na qual, de posse das informações geradas pela mineração de dados, o especialista no domínio abordado interpretará, com auxílio do engenheiro de conhecimento, os modelos e padrões gerados pelo processo (GOLDSCHIMIDDT; PASSOS, 2005).

Figura 2- Fases do Processo de Descoberta de Conhecimento



Fonte: FAYYAD apud OLIVEIRA, 2013

2.2.1.1 A atividade de mineração de dados

Segundo Fayyad (1996) a atividade de mineração de dados, participante da etapa de processamento, atua na utilização de algoritmos específicos na busca por padrões a partir dos dados, sendo dividida nas tarefas de modelagem de previsão, análise de associação, análise de agrupamento de dados e detecção de anomalias, a mineração contribui de maneira significativa ao KDD, dando suporte e tornando-se a principal atividade desse processo.

Usada na descoberta de padrões nos dados, a tarefa de análise de associação busca extrair características que possuam grande índice de associação dentro dos dados, seja na forma de regras ou subconjuntos de características (KUMAR et al, 2009). Um exemplo nesse sentido pode ser tomado por supermercados, quanto à compra de itens por seus clientes. Nesse contexto pode-se verificar, por exemplo, em qual horário determinados itens poderão ser mais comprados, desenhando, desta forma, o perfil de seu cliente naquele momento do dia, isso sob a associação dos itens comprados por meio de regras para extração de potenciais

oportunidades para venda de dois ou mais itens de uma única vez.

Já a análise de agrupamentos é uma tarefa ligada à mineração de dados, que procura estabelecer grupos, denominados de clusters, através de algoritmos que buscam alocar, em um mesmo grupo, registros que possuam características mais semelhantes do que registros que pertençam a outros grupos (KUMAR et al, 2009). Dessa maneira, recursos como medida de similaridade ou raciocínio probabilístico são utilizados em algoritmos de aprendizagem não-supervisionada como o k-means, k-medoids, x-means (MOORE;PELEG, 2000), dbscan entre outros na busca por uma melhor alocação dos registros de dados nos grupos, de acordo com seu padrão e características (HAN; KAMBER, 2006).

E por fim, a detecção de anomalias refere-se à tarefa de identificação de comportamento nos dados que diferem dos demais no conjunto de dados (KUMAR et al, 2009). A exemplo de tal tarefa pode-se citar a detecção em fraudes de cartão de crédito, onde a empresa mantém um banco com as transações realizadas por seus clientes. No momento que é realizada uma transação não corriqueira pelo usuário do cartão o algoritmo detecta esse comportamento anômalo e o registra para posterior verificação junto ao cliente pela empresa fornecedora do cartão.

2.2.2 Algoritmos para análise de agrupamentos

A análise de agrupamentos, também chamada de segmentação de dados por dividir em grupos os dados de grandes bases de acordo com sua similaridade, possui uma gama de algoritmos e técnicas para sua execução. Tais algoritmos são baseados em densidade dos grupos ou na distância entre os elementos do grupo e um centroide, elemento cujo qual cada grupo é representado e que possui uma medida de similaridade entre os elementos do mesmo (HAN; KAMBER, 2006).

Para criação dos grupos, algoritmos baseados em densidade, utilizam-se da medição da densidade do grupo no processo de agrupamento, enquanto que algoritmos baseados em distância utilizam-se de medidas de similaridade baseadas na distância entre os elementos de um grupo e seu centroide (KUMAR et al, 2009). Ainda para Kumar et al (2009) o protótipo do grupo, centroide, é ajustado a cada iteração do algoritmo procurando estabelecer da melhor maneira sua posição na tarefa de agrupamento.

Segundo Han (2006), dentre as medidas de similaridade pode-se citar a distância euclidiana, a distância Manhattan e a distância Minkowski. A distância euclidiana, mais popular métrica, é definida como:

$$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{in} - x_{jn}\right)^2} \quad (1)$$

Onde i = (x_{i1}, x_{i2},…, x_{in}) e j = (x_{j1}, x_{j2},…, x_{jn}) são dois objetos de dados n-dimensionais. Sendo essa métrica de similaridade a distância entre dois pontos no plano cartesiano.

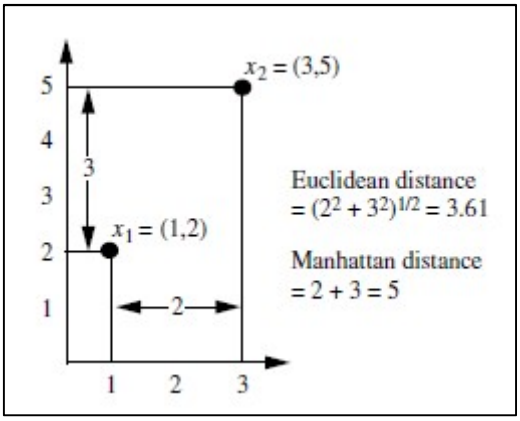
Em seguida, a distância Manhattan é uma métrica bastante conhecida e também utilizada no processo de agrupamento de dados (HAN; KAMBER, 2006), sendo esta definida por:

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right| \quad (2)$$

Onde i = (x_{i1}, x_{i2},…, x_{in}) e j = (x_{j1}, x_{j2},…, x_{jn}) são dois conjuntos de dados n-dimensionais.

Tal métrica faz alusão aos quarteirões da ilha de Manhattan em Nova Iorque, destacando-se como a distância entre dois pontos da ilha e que deve ser percorrida respeitando-se as suas ruas e quarteirões e não em linha reta como a distância euclidiana, conforme Figura 3.

Figura 3- Ilustrativo da distância Manhattan entre dois pontos



Fonte: HAN e KAMBER, 2006

E por fim, a generalização de ambas as distâncias supracitadas, euclidiana e Manhattan, é chamada distância Minkowski e é definida, segundo Han 2006 por:

$$d(i,j) = \sqrt[p]{\left|x_{i1} - x_{j1}\right|^p + \left|x_{i2} - x_{j2}\right|^p + \cdots + \left|x_{in} - x_{jn}\right|^p} \quad (3)$$

Onde *p* é um número inteiro positivo, $i = (x_{i1}, x_{i2},\ldots, x_{in})$ e $j = (x_{j1}, x_{j2},\ldots, x_{jn})$ são dois conjuntos de dados n-dimensionais. Esta distância também é conhecida como norma, onde *p=1* refere-se à distância Manhattan e *p=2* refere-se à distância euclidiana (HAN; KAMBER 2006). Interessante notar que *p* não deve ser confundido com o número de dimensões, atributos, dos dados em tela (KUMAR et al, 2009).

2.2.2.1 Uma técnica baseada em centroide: k-means

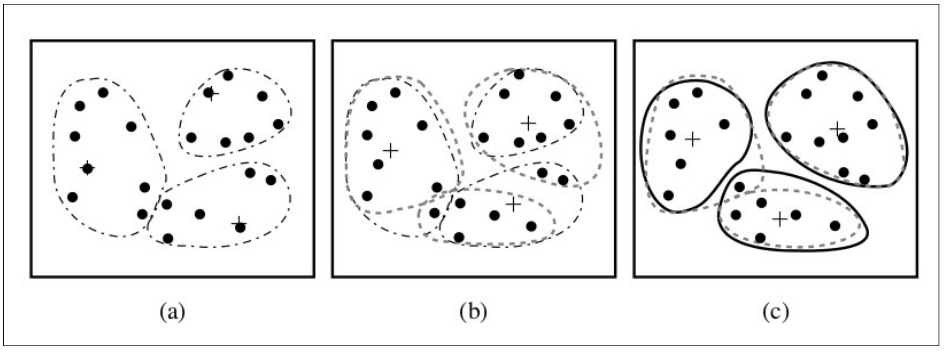
Muito utilizada atualmente, a técnica denominada k-means é baseada na busca pelo centroide dos grupos de acordo com a proximidade dos dados para com este centroide, utilizando o parâmetro de entrada *k*, informado pelo usuário, que diz respeito à quantidade de grupos da divisão dos dados (BERRY; LINOFF, 2004). Publicado em 1967 por J. B. MacQueen, o k-means é descrito em três passos por Berry e Linoff (2004):

i. Aleatoriamente o algoritmo aloca a quantidade de centroides informada pelo usuário através do valor de k, conforme Figura 4.a, que trata de uma ilustração da atuação do algoritmo para k = 3, onde os centroides são representados por + e os dados agrupados por ●.

ii. Em seguida o k-means associa cada registro ao seu centroide mais próximo definido, por meio de métrica anteriormente estabelecida, tal qual distância euclidiana ou distância Manhattan[11], conforme item b da Figura 4.

iii. A seguir novos centroides são calculados, através do valor médio dos pontos do grupo, para manterem-se o mais representativos possível dos dados que estão sendo agrupados, não sendo estes exatamente um elemento da massa de dados (HAN; KAMBER, 2006), tal qual encontra-se no item c Figura 4.

Figura 4 - Atuação do k-means para k=3



Fonte: HAN e KAMBER, 2006

[11] Definidas e mostradas na seção 2.2.2 do presente capítulo.

Tal processo é iterativo e finaliza-se quando não há mudanças significativas nos centroides definidos pelo passo 3, tornando-se estes os protótipos dos grupos definidos pelo processo de agrupamento de dados através do algoritmo k-means, que gerará grupos globulares visto que são tomados como centroides as médias dos valores dos pontos no grupo (KUMAR et al, 2009).

Diversas variações podem ser encontradas deste algoritmo, dentre elas pode-se citar o k-medoids, algoritmo cujo qual utiliza ao invés da média dos pontos do grupo para definir o centroide, o elemento mais representativo e que o denomina medoide (KUMAR et al, 2009), também pode-se citar o x-means, um poderoso algoritmo que propõe algumas melhorias frente o x-means (MOORE; PELEG, 2000) e que foi usado como ferramenta de agrupamento na presente dissertação.

2.2.2.2 X-means: uma extensão do k-means

Algoritmo desenvolvido por Moore e Pelleg no ano 2000, e publicado através do trabalho *"X-means: Extending K-means with Efficient Estimation of the Number of Clusters"*, o x-means propõe o processo de agrupamento de dados algumas melhorias em relação ao k-means, procurando dessa forma atingir um melhor resultado no processo de descoberta de conhecimento.

Apesar de muito utilizado e tendo a simplicidade de seu método como ponto positivo, o k-means possui também algumas limitações e fragilidades em sua atuação, sendo então prejudicada a sua execução em alguns casos específicos. Segundo Moore e Pelleg tais deficiência são:

i) Lentidão no tempo para completar cada iteração do algoritmo;

ii) O conhecimento do número cujo qual quer se dividir a base de dados, ou seja, a informação do valor de k deve ser inserida pelo usuário no processo de agrupamento, não sendo tão simples o fornecimento de tal dado ao algoritmo;

iii) Problema com convergência do algoritmo para mínimos locais. O algoritmo acredita que "encontrou" um bom modelo porém está preso em uma situação conhecida como mínimo local, ou seja, um modelo falsamente ideal no agrupamento de dados realizado pelo k-means (NORVIG; RUSSEL, 2004).

Assim, a fim de corrigir essas deficiências e melhorar o funcionamento do mesmo, o x-means toma como parâmetro de entrada um limite inferior e superior para os valores de k (quantidade de grupos do processo), além do conjunto de dados (D) que serão agrupados, e

propõe como solução uma nova técnica utilizando os pontos positivos e as potencialidades do k-means, enfrentando problemas existentes pelo mesmo (MOORE; PELLEG, 2000). Ainda segundo Moore e Pelleg (2000), o algoritmo utiliza-se de duas operações que se repetem até completarem o agrupamento dos dados, sendo então este definido nos três passos que se seguem:

1) Melhoria dos Parâmetros;

2) Melhoria da Estrutura;

3) Se k > kmax então pare e reporte o modelo de melhor pontuação (*score*) encontrado durante o processo de busca senão retorne para passo 1.

A fase 1, composta pela operação de melhoria de parâmetros, consiste simplesmente na execução do tradicional k-means para convergência do resultado. Enquanto a fase 2, denominada melhoria da estrutura, refere-se à verificação e necessidade do aparecimento de novos centroides nos dados em questão. Caso positivo, nessa fase há também a descoberta da localização desses novos protótipos (MOORE; PELLEG, 2000).

Segundo Moore e Pelleg (2000), o surgimento e estabelecimento dos novos centroides são verificados através de um recurso denominado Critério de Informação Bayesiano (Bayesian Information Criterion - BIC Scoring), que através da probabilidade a posteriori estabelece uma pontuação para o modelo encontrado e verifica se realmente é necessário o surgimento desses novos centros. Tal critério se dá pela equação de Kass e Wasserman (1995) que se segue:

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2}.\log R \qquad (4)$$

Onde, $\hat{l}_j(D)$ é a função de verossimilhança dos dados (D) de acordo com o j-ésimo modelo tomada no maior ponto de probabilidade. Sendo p_j a quantidade de parâmetros de M_j, a família de alternativas de modelos, que aqui correspondem aos diferentes modelos criados para diferentes valores de k, lembrando que como parâmetro de entrada do algoritmo é inserido o limite inferior e superior para a quantidade grupos que será estabelecida (MOORE; PELLEG, 2000).

Segundo Moore e Pelleg (2000), a máxima probabilidade estimada (MLE) para a variância, utilizando-se a esfera gaussiana, já que no k-means os grupos originados são globulares, se dá pela equação:

$$\hat{\sigma}^2 = \frac{1}{R-K}\sum_i \left(x_i - \mu_{(i)}\right)^2 \quad (5)$$

A probabilidade nos pontos é:

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} . \frac{1}{\sqrt{2\pi\hat{\sigma}^M}} . \exp\left(-\frac{1}{2\hat{\sigma}^2}\left|\left|x_i - \mu_{(i)}\right|\right|^2\right) \quad (6)$$

A verossimilhança dos dados é dada por:

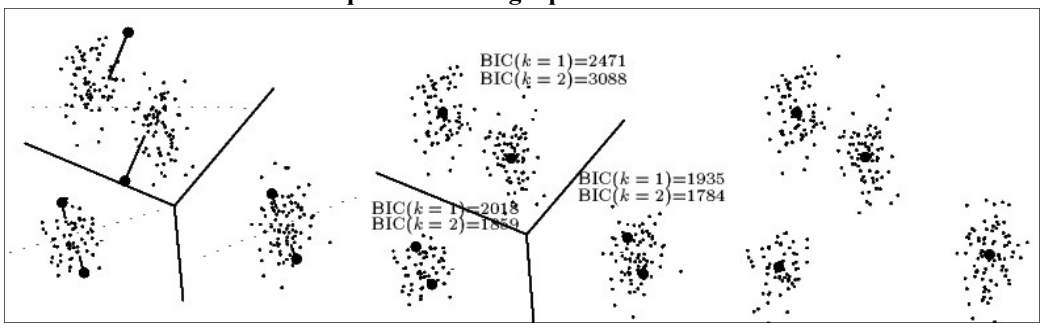
$$l(D) = log \prod_i P(x_i) = \sum_i \left(log \frac{1}{\sqrt{2\pi\sigma^M}} - \frac{1}{2\sigma^2}\left|\left|x_i - \mu_{(i)}\right|\right|^2 + log \frac{R_{(i)}}{R}\right) \quad (7)$$

E assim, deixando fixo n entre 1 e K, e focando no conjunto de pontos D_n que pertencem ao centroide n tem-se:

$$\hat{l}(D_n) = -\frac{R_n}{2}\log(2\pi) - \frac{R_n . M}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n log R_n - R_n log R \quad (8)$$

Logo, através do critério BIC é possível a escolha do melhor modelo no tocante a quantidade de grupos gerados pelo processo de mineração de dados. Dessa forma, a cada iteração a subdivisão dos centroides, respeitando-se o kmax informado como parâmetro de entrada do algoritmo, é avaliada calculando-se o BIC do novo estado, cujo qual é comparado com o BIC do modelo antes da divisão, prevalecendo assim o modelo com a maior pontuação para o critério (MOORE; PELLEG, 2000), conforme processo de agrupamento utilizando x-means e que pode ser visto na Figura 5.

Figura 5 - Visualização do BIC no processo de agrupamento utilizando o x-means para escolha da melhor quantidade de grupos nos dados



Fonte: MOORE e PELLEG, 2000

Dessa forma o problema 2 do algoritmo k-means, que diz respeito a quantidade de grupos informada antes da execução do algoritmo se resolve, já que esse valor será obtido pelo x-means através do uso do BIC.

Para correção da fragilidade 1, problemas com o tempo de execução do algoritmo, o x-means utiliza-se de uma *kd-tree* para incorporar os dados que serão utilizados em seus nós juntamente com estatísticas suficientes ao processo. Adicionalmente pode utilizar-se de uma blacklisting a fim de considerar apenas os centroides necessários à execução do algoritmo, permitindo assim um melhor desempenho do mesmo (MOORE; PELLEG, 2000).

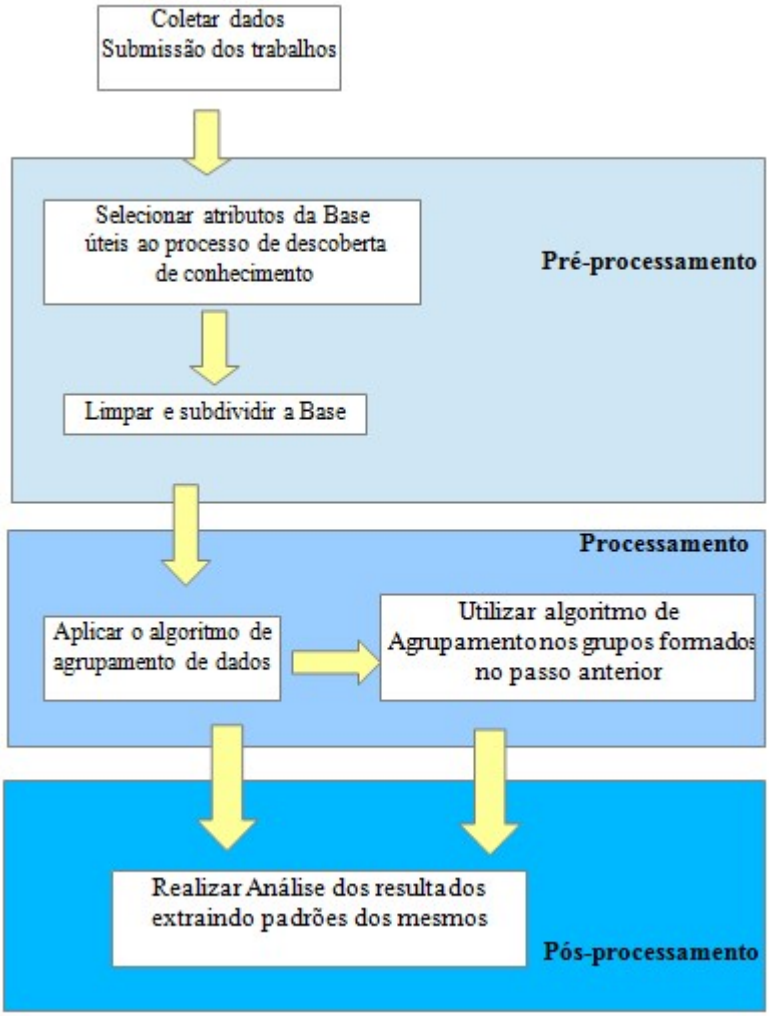
2.3 RapidMiner: Ferramenta Utilizada no Processo de Mineração de Dados

Na execução da tarefa de mineração de dados no presente trabalho, utilizou-se uma ferramenta cuja proposta é de otimizar o desenvolvimento e preparação dos experimentos realizados no processo de descoberta de conhecimento. Assim, o software RapidMiner em sua versão 5.2 foi utilizado, proporcionando agilidade e praticidade na etapa de mineração de dados. No apêndice A encontra-se a descrição do operador x-means implementado no RapidMiner e que foi utilizado no presente trabalho para agrupamento dos dados em grupos na etapa de mineração do KDD.

3 PROCESSO DE DESCOBERTA DE CONHECIMENTO QUANTO À PRODUÇÃO NO VI CONNEPI

O processo de Descoberta de Conhecimento, conforme vislumbrado no Capítulo 3, será adotado como base para o desenvolvimento deste capítulo. Tomando as etapas de Pré-processamento, Processamento e Pós-Processamento como caminho a ser traçado, espera-se que ao final seja possível a detecção do conhecimento implícito na base de dados. Com base nisso, é possível visualizar através de um fluxograma as fases que serão percorridas neste trabalho, conforme Figura 6.

Figura 6 - Fluxo do processo de descoberta de conhecimento aplicado ao presente trabalho



Fonte: Autor, 2013

3.1 Pré-processamento

Etapa muito importante para o desenvolvimento do Processo de Descoberta de Conhecimento, o pré-processamento trata desde a obtenção dos dados até sua limpeza e seleção. Tal etapa não pode ser desprezada, pois é o passo inicial no KDD, já que prepara os dados de entrada que alimentarão o algoritmo de aprendizagem de máquina na etapa de processamento e caso não seja realizada corretamente, poderá acarretar problemas futuros tanto na execução da etapa de processamento quanto na fase de pós-processamento, gerando uma análise equivocada a respeito da aprendizagem realizada.

3.1.1 Coleta de dados e seleção de atributos

Os dados utilizados para extração do conhecimento correspondem aos dados das avaliações dos artigos submetidos no VI CONNEPI, esses foram obtidos a partir da base de dados relacional, em PostgreSQL, gerada na ocasião do evento em 2011 e gerenciada através do sistema de submissões desenvolvido no Instituto Federal de Sergipe cuja finalidade foi realizar o cadastro dos pesquisadores, permitir o envio do artigo on-line pelo mesmo mediante login e senha, além de todo o gerenciamento das avaliações.

Após a coleta dos dados realizada pelo sistema de submissão, foi realizada uma seleção dos atributos úteis à pesquisa em tela. Por conta da grande quantidade de atributos e da complexidade do sistema, foi necessária a preparação de uma consulta sql para obtenção dos dados diretamente na base. Após a execução do *script* foram trazidos do banco de dados as informações dos atributos listados na Tabela 1.

Tabela 1- Atributos selecionados para coleta de dados da base

Número	Atributo	Descrição
1	id_avaliacao	Número de identificação da avaliação de determinado artigo
2	id_artigo	Número de identificação do artigo
3	tx_titulo	Título do artigo
4	tx_nome	Nome do autor
5	instituicao_autor	Nome da instituição do autor
6	in_relevancia_do_tema	Nota do quesito relevância do tema
7	in_qualidade_tecnico_cientifica	Nota do quesito qualidade técnico-científica
8	in_originalidade	Nota do quesito originalidade
9	in_apresentacao_e_estrutura	Nota do quesito apresentação e estrutura
10	in_fundamentacao_teorica	Nota do quesito fundamentação teórica
11	in_recomendacao_final	Recomendação final do avaliador
12	id_tipo_apresentacao	Recomendação do avaliador quanto ao tipo de apresentação do trabalho
13	sub_area_tematica	Sub-área temática do artigo
14	area_tematica	Área temática do artigo
15	tx_comentarios_para_autores	Comentário dos avaliadores para os autores
16	tx_comentarios_para_comite_tecnico_cientifico	Comentários dos avaliadores ao comitê técnico científico

Fonte: Autor, 2013

É de extrema importância ressaltar que os itens de 6 a 11 correspondem aos quesitos do questionário preenchido pelo avaliador acerca do artigo e que possuem como valores as notas de 1 a 4, conforme Tabela 2.

Tabela 2 - Itens de respostas para atributos 7 a 11 da Tabela 1

Nota	Conceito
1	Ruim
2	Regular
3	Bom
4	Ótimo

Fonte: Autor, 2013

O item 11 da Tabela 1 trata da decisão final do avaliador a respeito do artigo cujas opções disponíveis ao mesmo encontram-se na Tabela 3.

Tabela 3 - Itens de respostas para o atributo 11

Nota	Conceito
1	Rejeitado
2	Aceito com restrições
3	Aceito sem restrições

Fonte: Autor, 2013

3.1.2 A limpeza e subdivisão dos dados utilizados

Na sequência, após a extração, foram retiradas todas as avaliações de artigos que não fossem de Institutos Federais, além das avaliações que possuíam como recomendação final o status 2. Pois tal nota indica que o artigo foi devolvido ao autor para correções, restando apenas algumas modificações a serem feitas. Já constando nos dados, com as mesmas notas para os itens, a avaliação, porém com status 3 para recomendação final, ou seja, a nota de aprovação do avaliador após ler as alterações feitas pelos autores.

Logo após a eliminação das avaliações com status 2 para recomendação final, foi realizada a divisão dos dados em dois grupos: avaliações de artigos do Instituto Federal de Sergipe e avaliações de artigos dos outros institutos federais. Tal divisão se fez necessária por conta da intenção de realizar uma avaliação comparativa do IFS para com os outros institutos do país, averiguando assim sua situação no cenário nacional. Desse processo resultou uma base de 4.665 avaliações de artigos de pesquisadores de todo o país e outra com 187 avaliações de artigos de pesquisadores apenas do Instituto Federal de Sergipe.

3.2 Processamento

Na etapa de processamento foi executada a mineração de dados em si. Dividido em dois momentos, o processamento contou com a utilização do algoritmo de aprendizagem não-supervisionada para identificação das características dos artigos/pesquisas contidas nas bases, agora divididas em duas conforme etapa de subdivisão dos dados acima citadas. Aqui foram usadas as tarefas de descrição para que fosse possível subsidiar a etapa seguinte, o pós-processamento, na realização das análises acerca da obtenção do conhecimento implícito nos dados obtidos. Para isso o software RapidMiner[12] foi utilizado, agilizando assim aplicação dos algoritmos já implantados em seu núcleo (core).

[12] Observar item 2.3 do capítulo 2.

3.2.1 Algoritmo de agrupamento de dados para geração de grupos

Nesse primeiro momento do processamento, os dados foram submetidos ao algoritmo de agrupamento chamado x-means[13]. Trata-se de uma variação do k-means[14], popular algoritmo que cria protótipos em meio aos dados dos grupos construídos. Tais protótipos, chamados de centroides, servem para aglomerar os dados que possuem características semelhantes a sua volta, proporcionando assim divisão em grupos com propriedades congruentes.

Dentre as vantagens do x-means apresentadas no capítulo 2, foi buscada neste presente trabalho, a obtenção automática da quantidade de grupos em que devem se dividir os dados. Assim, o valor de k, antes informado ao algoritmo de agrupamento de dados k-means, foi automaticamente definido pelo processo de mineração, estabelecendo dessa forma a sua divisão natural.

Durante essa fase, dois processos distintos foram executados, os quais chamamos de Experimento 1 e Experimento 2. No primeiro experimento foi aplicado o algoritmo x-means para subdivisão dos dados em grupos distintos. Enquanto que no segundo experimento um grupo gerado no Experimento 1, mereceu mais estudo e esse foi submetido ao algoritmo x-means, com o objetivo de entende-lo melhor. Em ambos os casos as duas bases geradas na etapa de pré-processamento foram utilizadas como entrada para os algoritmos de mineração de dados, ou seja, os agrupamentos criados utilizando-se a base de avaliações de artigos do IFS e dos demais Institutos Federais de Educação.

3.2.1.1 Experimento 1

Como mencionado anteriormente, aqui os dados foram submetidos ao x-means cuja configuração realizada pode ser observada na Tabela 4.

---

[13] Descrito no item 2.2.2.2 do capítulo 2.
[14] Descrito no item 2.2.2.1 do capítulo 2.

Tabela 4 - Configuração do algoritmo x-means para Experimento 1

Atributo	Valor
K min	2
K max	10
Tipo de medição de distância	Medição Numérica
Métrica	Distância Euclidiana

Fonte: Autor, 2013

Primeiramente para esse experimento, os dados das avaliações dos artigos de todos os Institutos Federais exceto o IFS foram submetidos ao processo de aprendizagem não-supervisionada, nomeando assim este como Experimento 1-A. Logo após, foi a vez dos dados de avaliações dos artigos do Instituto Federal de Sergipe passarem pelo processo de agrupamento utilizando-se mesma configuração para o x-means aplicada na base anterior, conforme Tabela 4, denominando-se este como Experimento 1-B.

3.2.1.2 Experimento 2

Nesse experimento, utilizou-se novamente o algoritmo de agrupamento x-means. Dessa vez utilizado para obtenção automática dos subgrupos de um grupo em específico, o grupo de Desempenho Heterogêneo (DH), do Experimento 1. Tais subgrupos serviram para estudo das características e comportamento dos artigos quanto a aceitação ou rejeição no congresso alocados nesse grupo de desempenho heterogêneo. Aqui o algoritmo recebeu como parâmetros de configuração os mesmos utilizados no Experimento 1, conforme Tabela 4.

O experimento contou com execução em dois momentos, assim como o Experimento 1, primeiramente o grupo DH originado pelos dados de avaliação dos artigos de todos os institutos federais à exceção do IFS foram submetidos ao algoritmo x-means, denominando-se Experimento 2-A. Em seguida as avaliações dos pesquisadores do Instituto Federal de Sergipe foram submetidas ao mesmo processo, nomeando este como Experimento 2-B, gerando assim subgrupos no grupo DH já obtido no Experimento 1.

Importante ressaltar que em ambos os experimentos, 1 e 2, espera-se obter os dados categorizados nos devidos grupos, ou seja, espera-se que o algoritmo x-means agrupe os dados com um novo atributo chamado cluster, realizando assim a subdivisão dos mesmos em grupos de acordo com as relações entre as suas características detectadas no processo de mineração de dados pelo algoritmo em questão.

3.3 Análise dos Resultados para Extração de Características e Avaliação da Solução Proposta

De posse dos grupos criados no Experimento 1 e dos subgrupos originados no Experimento 2 pelo algoritmo de agrupamento foi possível estabelecer relações entre os atributos e definir características das submissões dos pesquisadores dos institutos federais do país no VI CONNEPI. Nesse momento também foi realizada uma análise que expôs e trouxe à compreensão uma comparação entre essas caraterísticas das submissões dos pesquisadores do IFS em contraste às características das submissões dos demais pesquisadores dos institutos federais.

Após isso puderam ser detectadas tanto fragilidades quanto potencialidades dos pesquisadores do Instituto Federal de Sergipe, informações cujas quais a Pró-reitoria de Pesquisa e Extensão poderá se valer em seu processo de tomada de decisão, para assim fortalecer a atividade de pesquisa no instituto.

4 RESULTADOS E ANÁLISES DOS AGRUPAMENTOS

Inicialmente, o presente capítulo procura trazer informações sobre as bases de dados utilizadas no processo de descoberta de conhecimento, além de descrever e exibir os resultados obtidos nos experimentos definidos no Capítulo 3. Tais resultados, sejam eles em forma de tabelas ou gráficos, são de grande importância na busca pelo alcance dos objetivos propostos no Capítulo 1 do presente trabalho.

4.1 Caracterização dos Dados Utilizados

Os dados utilizados no processo foram obtidos, através das submissões e avaliações dos trabalhos no VI CONNEPI[15]. Durante a etapa de pré-processamento[16], os dados tiveram que passar por alguns procedimentos de limpeza para que assim pudessem ser utilizados adequadamente neste trabalho.

Primeiramente, a extração se deu através de um script escrito em SQL (Structured Query Language) que originou desta maneira as duas bases seguintes:

- Base com dados de avaliações de todos os Institutos exceto o IFS;
- Base com dados de avaliações do IFS;

Inicialmente as bases acima citadas possuíam registros de avaliações que retornaram ao autor para correção. Logo após isso, tais registros foram removidos para que não atrapalhassem o processo de mineração, já que os mesmos em sua grande maioria estariam já inseridos na base, porém com o item “Recomendação Final” valorado como 3 (Aprovado), o que duplicava sua ocorrência em meio aos dados a serem trabalhados.

Dessa maneira, após essa retirada, as bases apresentadas à mineração contêm o quantitativo de registros conforme visualizado na Tabela 5.

---

[15] Conforme metodologia proposta no Capítulo 3.
[16] Trazida no item 3.1 no Capítulo 3.

Tabela 5 - Quantitativo de registros nas bases de dados após etapa de pré-processamento

Base	Quantidade de Registros
Dados de Todos os Institutos exceto IFS	4.665
Dados do IFS	187

Fonte: Autor, 2013

Outra importante providência tomada, para facilitar o entendimento e execução da leitura de gráficos e outras informações, é que a partir de agora, os itens utilizados na avaliação dos artigos encontram-se juntamente com sua devida abreviação, conforme Tabela 6.

Tabela 6 - Identificação dos itens de avaliação dos artigos

Item	Abreviação
Relevância do Tema	RT
Qualidade Técnico Científica	QTC
Originalidade	OR
Apresentação e Estrutura	AE
Fundamentação Teórica	FT
Recomendação Final	RF

Fonte: Autor, 2013

4.1.1 Caracterização da base de dados com avaliações de todos os institutos federais de educação ciência e tecnologia à exceção do IFS

De posse dos dados a serem utilizados, é possível observar suas características e obter assim importantes informações. Uma delas é o comportamento das respostas dos avaliadores nos mais diversos itens de avaliação dos artigos. Conforme Tabela 7, percebe-se notoriamente a presença marcante da resposta "*Bom*" em uma maior quantidade de artigos em seus diversos itens de avaliação, ocupando 54,33% das respostas dos avaliadores.

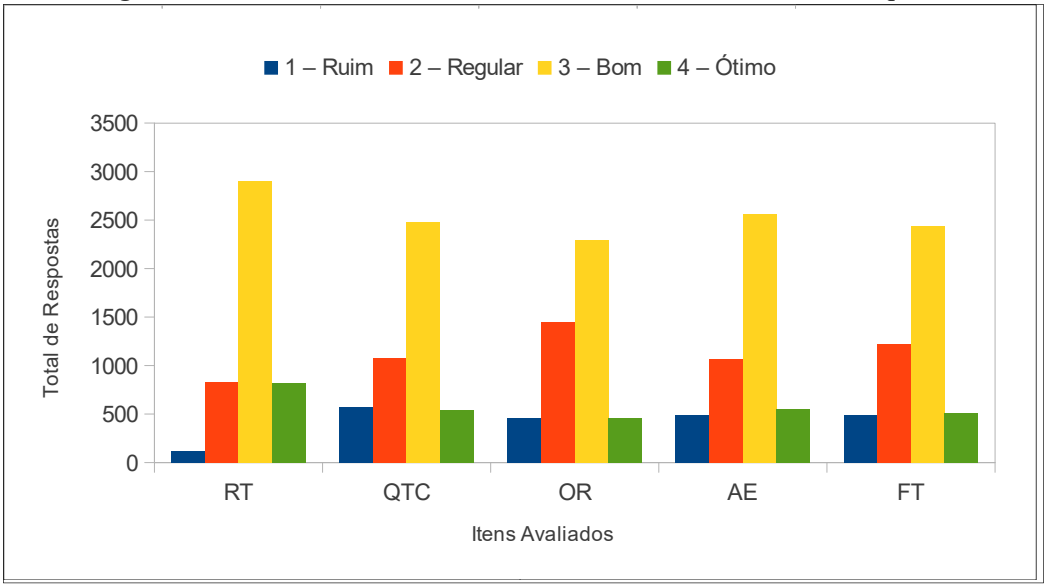
Tabela 7 - Dados tabulados das respostas dos avaliadores por item

Opção	RT	QTC	OR	AE	FT	Frequência	Freq. (%)
1 - Ruim	122	571	463	489	489	2134	9,15
2 - Regular	826	1081	1452	1065	1221	5645	24,20
3 - Bom	2901	2475	2292	2563	2441	12672	54,33
4 - Ótimo	816	538	458	548	514	2874	12,32

Fonte: Autor, 2013

Outra importante ferramenta descritiva para caracterização dos dados é o histograma baseado na tabela do quantitativo das respostas, conforme Gráfico 8. Nele percebe-se um desempenho satisfatório do item "*Relevância do Tema*" (RT) em relação às demais pontuações dos outros itens de avaliação. Pois, além do mesmo obter poucas notas "*Ruim*" comparado com as demais questões, ele ainda manteve a maior quantidade de notas "*Bom*" em meio a todos os outros itens avaliados. Tal informação pode ser validada através de dados da Tabela 8, observando a maior média entre os itens e o menor desvio padrão entre os mesmos, caracterizando assim esse o melhor item de desempenho dos pesquisadores dos institutos de todo o país.

Gráfico 8 - Histograma dos dados de todos os institutos exceto o IFS em relação aos quesitos de avaliação



Fonte: Autor, 2013

Tabela 8 - Média e desvio padrão dos dados da base de avaliações de todos os institutos exceto o IFS

Item	Média	Desvio Padrão	Coeficiente de Dispersão (%)
RT	2,9	0,7	24,14
QTC	2,6	0,8	30,77
OR	2,6	0,8	30,77
AE	2,7	0,8	29,63
FT	2,6	0,8	30,77

Fonte: Autor, 2013

Para finalizar a caracterização dos dados das avaliações de todos os institutos a exceção do IFS, é possível verificar sua média e dispersão nos itens julgados. Pode-se

visualizar a tendência já apontada pelo Gráfico 8, dos dados manterem-se em torno da nota 3 (“*Bom*”) para os itens avaliados, com uma baixa dispersão, tendendo a concentrar assim os resultados em torno da média, sem grande dispersão de valores.

4.1.2 Caracterização da base de dados com avaliações de artigos do IFS

Trabalhando da mesma forma com os dados das avaliações do IFS, é possível perceber o comportamento das avaliações através do quantitativo de respostas aos itens avaliados no congresso. De acordo com a *Tabela 9* visualiza-se que o desempenho dos pesquisadores do Instituto Federal de Sergipe foi similar ao demonstrado pelos pesquisadores de todo o país, pois trouxe em sua maioria a resposta “*Bom*” aos itens avaliados. Mantendo a mesma tendência da base anterior para todas as outras respostas às questões.

Tabela 9 - Dados tabulados das respostas dos avaliadores por item

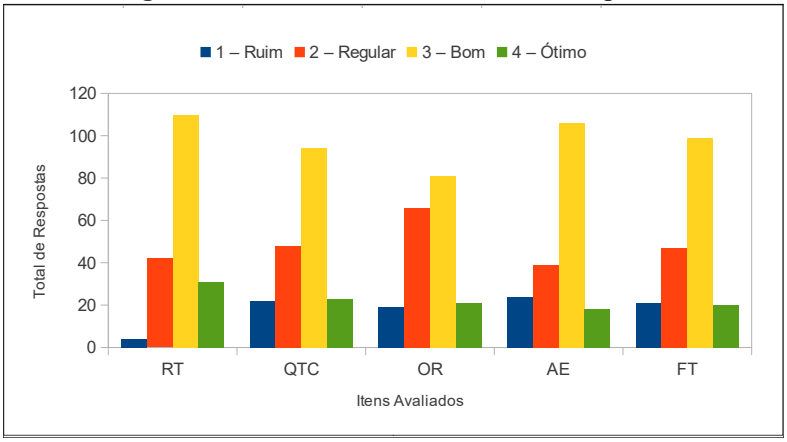
Opção	RT	QTC	OR	AE	FT	Frequência	Frequência (%)
1 - Ruim	4	22	19	24	21	90	9,63
2 - Regular	42	48	66	39	47	242	25,88
3 - Bom	110	94	81	106	99	490	52,41
4 - Ótimo	31	23	21	18	20	113	12,09

Fonte: Autor, 2013

No que diz respeito ao item de avaliação de melhor desempenho pelos pesquisadores do IFS, visualiza-se no Gráfico 9 que o quesito “*Relevância do Tema*” obteve um desempenho mais satisfatório, pois alcançou um grau mais elevado em sua resposta, mantendo a média, conforme Tabela 10, mais alta do que os demais itens, além de possuir o menor coeficiente de variação de todas as questões avaliadas, concedendo assim a esse uma dispersão consideravelmente satisfatória.

Além disso, o desempenho obtido pelos pesquisadores do IFS não está muito aquém do resultado alcançado pelos demais pesquisadores do país. Prova disto é que no mesmo item de menor desempenho para os pesquisadores do Brasil, o pesquisador do Instituto Federal também obteve um baixo desempenho. Como pode ser visto no Gráfico 9, o item de avaliação “*Originalidade*” refletiu nas avaliações dos pesquisadores do Instituto Federal de Sergipe o que ocorreu com pesquisadores de todo o país, uma nota mais baixa, que pode ser vislumbrada por meio de sua média conforme Tabela 10.

Gráfico 9 - Histograma dos dados do IFS em relação aos quesitos de avaliação



Fonte: Autor, 2013

Tabela 10 - Média e desvio padrão dos dados da Base do IFS

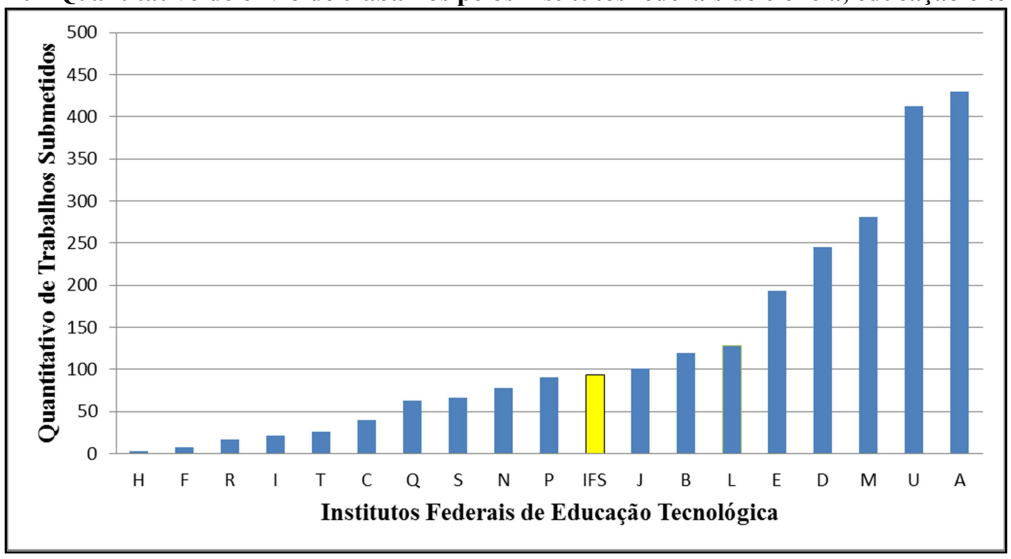
Item	Média	Desvio Padrão	Coeficiente de Variação (%)
RT	2,9	0,7	24,14
QTC	2,6	0,8	30,77
OR	2,6	0,8	30,77
AE	2,6	0,8	29,63
FT	2,6	0,8	30,77

Fonte: Autor, 2013

Além dos itens supracitados, o padrão de respostas também se mantém nos demais quesitos avaliados no congresso (Qualidade Técnico Científica, Apresentação e Estrutura, Fundamentação Teórica). Mostrando assim que o pesquisador do Instituto Federal de Sergipe possui semelhança, pelo menos nessa primeira análise, aos padrões avaliativos estabelecidos no congresso.

Assim, após a visualização dessas informações, é fato que o IFS mantém uma aproximação no resultado das avaliações realizadas dos artigos dos demais institutos federais no país, ocupando a 9ª colocação em meio aos 19 Institutos Federais de Educação Tecnológica no quantitativo de artigos enviados ao congresso. Estando também bem colocado no que diz respeito ao quantitativo de submissões ao congresso, conforme Gráfico 10.

Gráfico 10 - Quantitativo de envio de trabalhos pelos institutos federais de ciência, educação e tecnologia



Fonte: Autor, 2015

4.2 Resultados dos Experimentos Realizados

Aplicado em dois momentos distintos, o desenvolvimento do trabalho fez uso de um algoritmo de agrupamento de dados, o x-means, na descoberta de conhecimento. Sendo então estabelecido seu uso nas bases anteriormente separadas, e sua metodologia se deu por conta dos seguintes experimentos:

- Experimento 1: Agrupamento por meio do x-means.
- Experimento 2: Agrupamento por meio do x-means dos grupos heterogêneos quanto a aceitação dos artigos, gerados pelo Experimento 1. Esse experimento foi planejado após a análise do primeiro e a necessidade encontrada de se detalhar um pouco mais os artigos com esse perfil.

4.2.1 Experimento 1: agrupamento por meio do x-means

O experimento utilizou como algoritmo de clustering o x-means. Tal opção de utilização se deu devido ao fato do mesmo determinar a quantidade de grupos que os dados seriam divididos, por meio do recurso denominado BIC[17] (Bayesian Information Criterion). Dessa forma a tarefa de mineração de dados foi demasiadamente facilitada por conta dessa característica do algoritmo.

Assim, o Experimento 1 subdividido em dois, o experimento 1-A e o experimento 1-

[17] Conforme seção 2.2.2.2 do Capítulo 2.

B, as duas bases de dados de avaliações de artigos foram submetidas ao algoritmo de agrupamento de dados x-means. Focando dessa forma em encontrar relações futuras entre os dois experimentos realizados cujos resultados são vislumbrados nas seções que se seguem.

4.2.1.1 Experimento 1-A: aprendizagem baseada na base de dados de avaliações de todos os institutos federais de educação ciência e tecnologia à exceção do IFS

Este experimento utilizou-se da base de dados de avaliações de artigos de todos os institutos exceto o IFS para aplicação do x-means. Desse experimento obteve-se uma tabela com os centroides para os grupos criados automaticamente no processo de agrupamento, e que podem ser visualizados na Tabela 12. Uma importante observação que deve ser feita no tocante a esse experimento é que o mesmo originou a formação de quatro grupos distintos no processo de mineração de dados.

A saber tais grupos foram nomeados de acordo com o perfil das notas que abrange, e que pode ser visto na Tabela 11.

Tabela 11- Denominação dos grupos formados após mineração

Sigla	Nome
DR	Desempenho Ruim
DH	Desempenho Heterogêneo
DA	Desempenho Alto
DB	Desempenho Bom

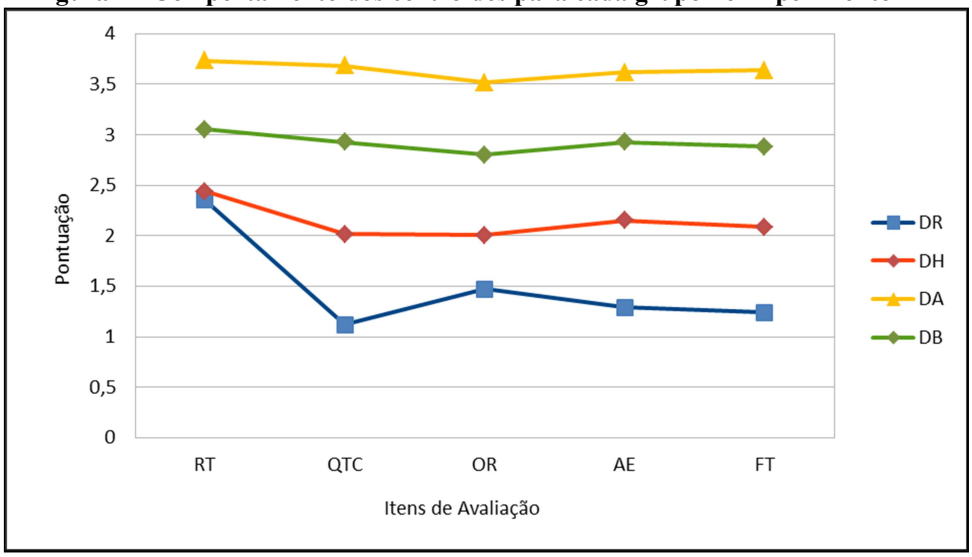
Fonte: Autor, 2013

Tabela 12 - Centroides dos grupos originados pelo x-means para a base de avaliação de todos os institutos exceto o IFS

Grupo	RT	QTC	OR	AE	FT	Total Itens (%)
DR	2,352	1,121	1,471	1,292	1,240	11,51
DH	2,441	2,015	2,006	2,151	2,088	21,35
DA	3,734	3,681	3,518	3,617	3,637	15,24
DB	3,053	2,925	2,802	2,929	2,882	51,89

Fonte: Autor, 2013

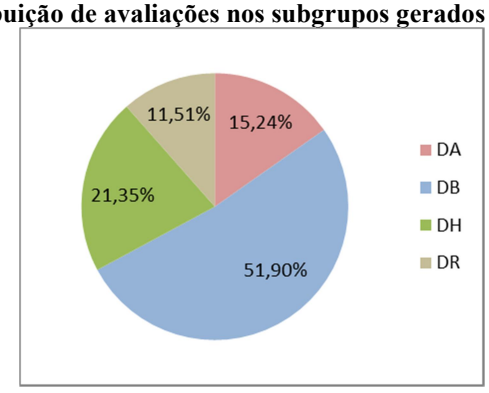
Figura 7 - Comportamento dos centroides para cada grupo no Experimento 1-A



Fonte: Autor, 2013

Dos quatro grupos apresentados por este método de agrupamento os dados ficaram distribuídos em seu interior e guiados pelos centroides da Tabela 12 da seguinte forma: DR com 537 itens (11,51%), DH com 996 (21,35 %), DA com 711 (15,24%) e DB com 2421 itens (51,89%), conforme Gráfico 11.

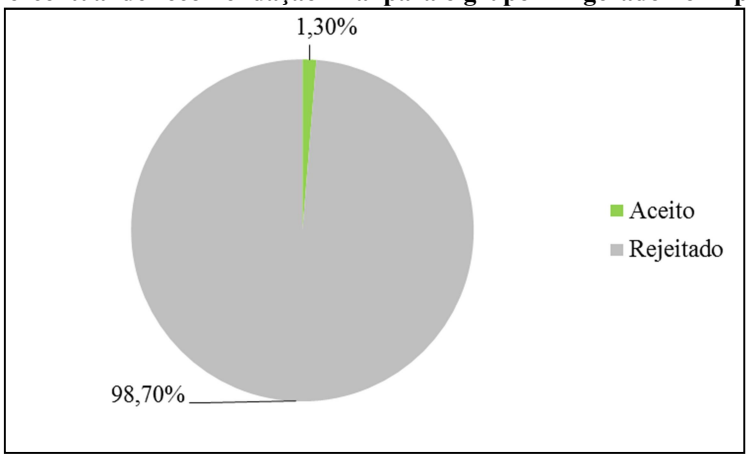
Gráfico 11 - Distribuição de avaliações nos subgrupos gerados no Experimento 1-A



Fonte: Autor, 2013

Pôde-se notar também, conforme Gráfico 12, no que diz respeito ao gráfico segundo a recomendação final da avaliação dos artigos para cada um dos quatro centroides dos grupos originados no processo, que o grupo Desempenho Ruim possui em sua maioria avaliações de artigos que foram rejeitados no processo de submissão e análise, pois seu centroide possui baixos valores para os atributos utilizados no agrupamento dos dados. Tal condição condiz com o atributo Recomendação Final observado na base de dados para esse grupo gerado.

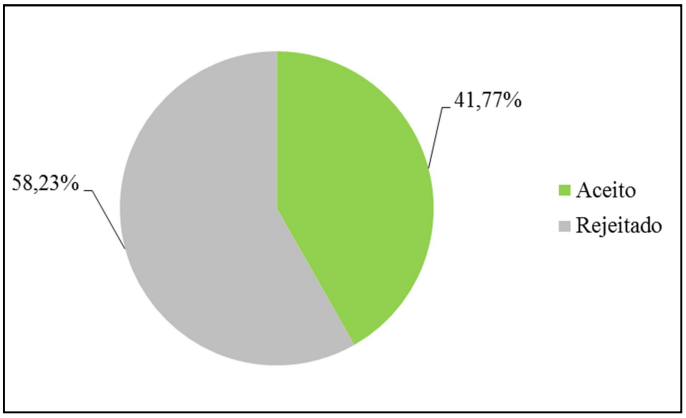
Gráfico 12 - Percentual de recomendação final para o grupo DR gerado no Experimento 1-A



Fonte: Autor, 2013

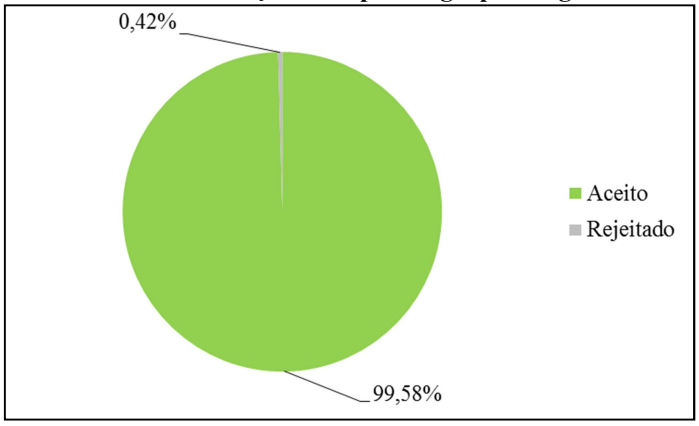
Por sua vez os demais grupos também guardam em seu interior dados de submissões que foram aceitas, conforme Gráfico 13, Gráfico 14 e Gráfico 15. Sendo o grupo Desempenho Alto o grupo com os trabalhos com as maiores notas adquiridas nas avaliações no congresso.

Gráfico 13 - Percentual de recomendação final para o grupo DH gerado no Experimento 1-A



Fonte: Autor, 2013

Gráfico 14 - Percentual de recomendação final para o grupo DA gerado no Experimento 1-A

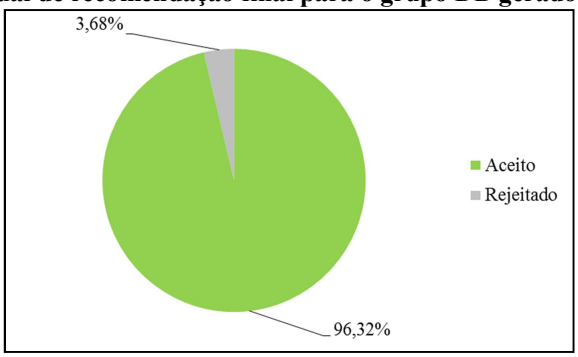


Fonte: Autor, 2013

Através dos grupos formados pelo algoritmo também é possível identificar características em comum entre as avaliações realizadas, tais como, para pertencer ao grupo DR o artigo deve ter obtido notas baixas nos itens avaliados, sendo a Relevância do Tema o item com maior pontuação entre os demais itens desse mesmo grupo e sendo o item Qualidade Técnico Científica o item de menor pontuação.

Para o grupo DH o comportamento do item Referencial Teórico mantém a mesma tendência encontrada no grupo DR, ele possui maior valor em média comparado aos demais itens de avaliação nesse grupo. Pode-se perceber também que os demais atributos mantêm-se mais constantes do que o grupo DR, não possuindo tanto decaimento do item Qualidade Técnico Científica perante os demais. Além do que o item Apresentação e Estrutura possui melhor nota do que o item Originalidade, justamente o oposto do que ocorre no grupo DR, no qual o item Originalidade sobressaiu sobre os demais itens a exceção da Relevância do Tema.

Gráfico 15 - Percentual de recomendação final para o grupo DB gerado no Experimento 1-A



Fonte: Autor, 2013

Concentrando 51,89% das avaliações realizadas, o grupo DB alocou a maioria dos trabalhos enviados ao congresso e também manteve o comportamento similar ao grupo DH no que diz respeito a tendência constante de suas notas. Inclusive tanto o atributo Originalidade quanto Apresentação e Estrutura mantiveram a tendência observada no grupo anterior, a de manter uma nota melhor para Apresentação e Estrutura perante a Originalidade.

Por fim, no grupo DA estão localizadas as maiores notas e os melhores desempenhos dos artigos submetidos ao evento, conforme Figura 7. Nesse grupo, que contém 15,24% das avaliações dos trabalhos submetidos, todos os atributos mantém uma diferença mínima do item Relevância do Tema, mas reflete o mesmo comportamento dos grupos DH e DB no que diz respeito à tendência dos itens Originalidade e Apresentação e Estrutura.

4.2.1.2 Experimento 1-B: aprendizagem baseada na base de dados de avaliações do IFS

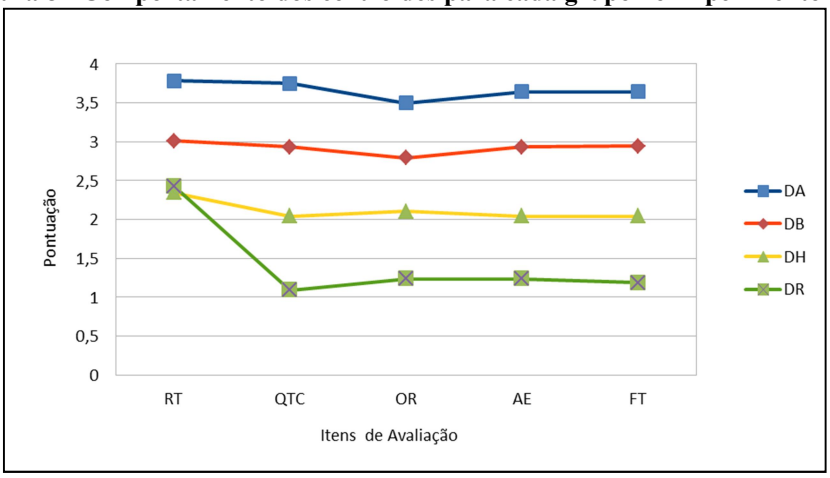
Nesse experimento foi utilizada como entrada ao algoritmo de agrupamento x-means a base de dados de avaliações de artigos do IFS. Seguindo o mesmo procedimento que o Experimento 1-A, o x-means através do Critério de Informação Bayesiana alocou os dados novamente em quatro grupos cujos quais possuem os centroides da Tabela 13. Utilizou-se a mesma nomenclatura do Experimento 1-A para os grupos originados a partir do processo de agrupamento de dados.

Tabela 13 - Centroides dos clusters originados pelo x-means para a base de avaliação de artigos do IFS no experimento 1-B

Grupo	RT	QTC	OR	AE	FT	Total Itens (%)
DA	3,785	3,75	3,5	3,642	3,642	14,97
DB	3,010	2,934	2,793	2,934	2,945	49,20
DH	2,347	2,043	2,108	2,043	2,043	24,60
DR	2,428	1,095	1,238	1,238	1,190	11,23

Fonte: Autor, 2013

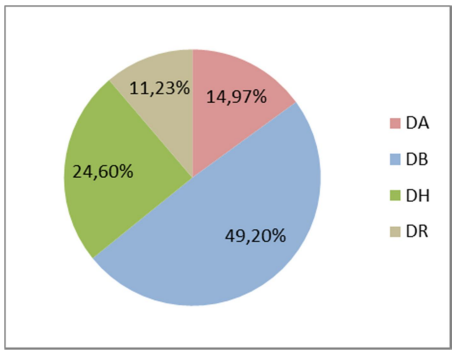
Figura 8 - Comportamento dos centroides para cada grupo no Experimento 1-B



Fonte: Autor, 2013

Dos quatro grupos apresentados pelo algoritmo de agrupamento, as avaliações dos artigos foram distribuídas em seu interior e guiadas pelos centroides, conforme Tabela 13, da seguinte forma: DA com 28 itens (14,97%), DB com 92 itens (49,20 %), DH com 46 itens (24,60%) e DR com 21 itens (11,23%), de acordo com Gráfico 16.

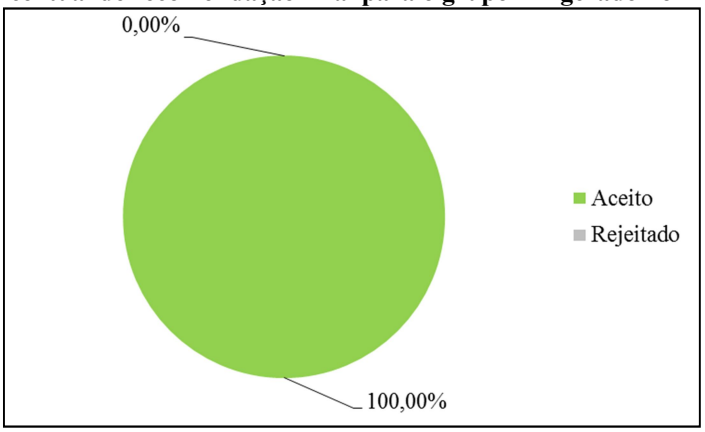
Gráfico 16- Distribuição de avaliações nos subgrupos gerados no Experimento 1-B



Fonte: Autor, 2015

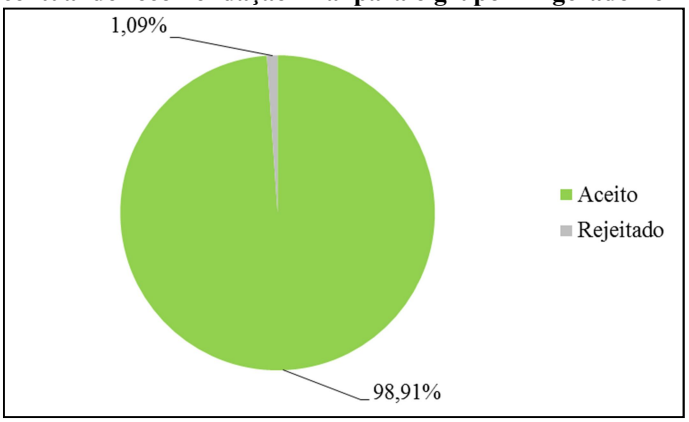
Como observado, conforme Gráfico 17, Gráfico 18, Gráfico 19 e Gráfico 20, enquanto o grupo DA concentra os dados de maiores notas nas avaliações os demais grupos gradativamente vão diminuindo suas notas até o alcance do grupo DR, que possui os dados de artigos rejeitados nas avaliações.

Gráfico 17 - Percentual de recomendação final para o grupo DA gerado no Experimento 1-B



Fonte: Autor, 2013

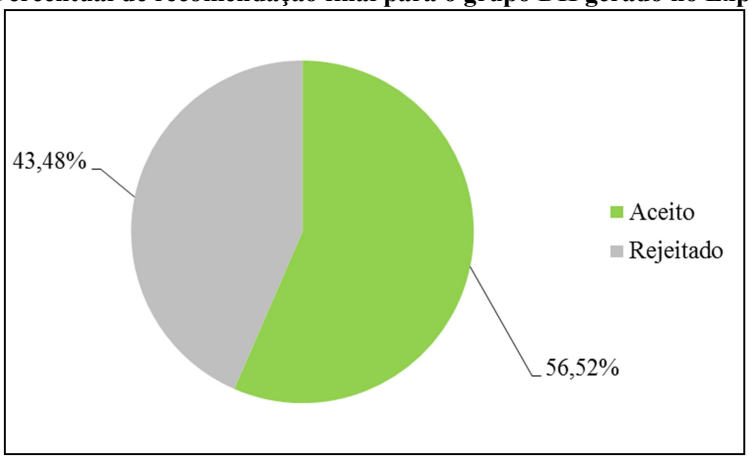
Gráfico 18 - Percentual de recomendação final para o grupo DB gerado no Experimento 1-B



Fonte: Autor, 2013

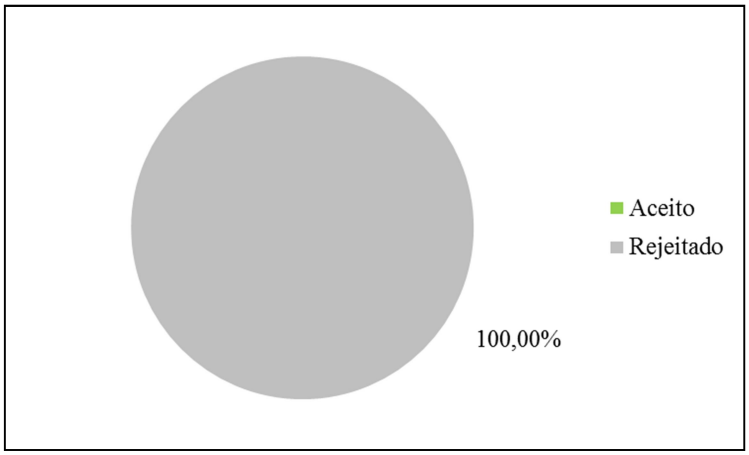
Um fato curioso pode ser notado apenas com a observação dos centroides encontrados pelo x-means e que podem ser vislumbrados na Tabela 13. A nota do atributo Relevância do Tema novamente possui o efeito do Experimento 1-A, ela é a maior dentre os outros atributos, porém neste experimento, esse comportamento se inverte nos grupos denominados DH e DR. Apesar do grupo DH possuir em média notas mais altas para todos itens de avaliação do que o grupo DR, nesse item em especial, Relevância do Tema, esse papel se inverte, cabendo ao grupo DR ser o possuidor da nota mais alta neste quesito.

Gráfico 19 - Percentual de recomendação final para o grupo DH gerado no Experimento 1-B



Fonte: Autor, 2013

Gráfico 20 - Percentual de recomendação final para o grupo DR gerado no Experimento 1-B



Fonte: Autor, 2013

Iniciando a análise pelo grupo DA, pode-se verificar que esse alocou 14,97% dos dados das avaliações de artigos do IFS no congresso, sendo tais avaliações as com maiores pontuações e melhor desempenho em todos os itens. Percebe-se também, através da Figura 8, que o item Originalidade obteve a menor pontuação dentre os demais, e que o item

Apresentação e Estrutura e o item Fundamentação Teórica mantiveram um equilíbrio em suas notas, ficando abaixo apenas do item Qualidade Técnico Científica, além claro do item Relevância do Tema.

Seguindo a mesma tendência do grupo DA o agrupamento denominado DB distingue-se apenas do primeiro pelo valor um pouco mais baixo de suas notas, mas mantém o mesmo relacionamento entre os seus atributos no que diz respeito ao equilíbrio e a tendência de suas notas.

Já nos grupos denominados DH e DR, observa-se um comportamento inverso quanto aos atributos Qualidade Técnico Científica, Originalidade e Apresentação e Estrutura. Enquanto que nos agrupamentos anteriores o item Originalidade estava sempre abaixo dos outros dois itens, aqui ele sobressai, possuindo valores acima dos outros dois atributos. Algo a perceber também no grupo DR, o cluster com maior quantidade de artigos rejeitados, que apesar dessa situação os artigos não possuem pontuação tão baixa em Relevância do Tema quanto os demais itens, provocando assim essa rampa acentuada na passagem entre tal item e o item Qualidade Técnico Científica, que é visualizada na Figura 8.

4.2.1.3 Comparativo entre os Experimentos 1-A e 1-B realizados

Com base nos experimentos realizados pode-se concluir a respeito do comportamento dos grupos criados pelo x-means o seguinte:

1) O item Relevância do Tema tem uma característica peculiar no conjunto de itens de avaliação, pois este possui as maiores notas de avaliação nos experimentos realizados.

2) Os grupos com maiores pontuações nos itens de avaliação possuem comportamento semelhante no tocante a seus atributos. Como por exemplo o comportamento do item Originalidade em ambos os experimentos o mesmo possui a característica de ter valores menores com relação aos demais itens para os grupos com pontuações geralmente mais baixas para os itens avaliados.

4.2.2 Experimento 2: agrupamento por meio do x-means dos grupos de desempenho heterogêneo obtidos no Experimento 1

Para este experimento aplicou-se o mesmo algoritmo, x-means, para os grupos de Desempenho Heterogêneo oriundos das bases já agrupadas do Experimento 1, este apresentado na seção 4.2.1 do presente capítulo. Assim, procurou-se elucidar o

comportamento das avaliações dos artigos dentro desse grupo no que diz respeito a aceitação ou rejeição dos mesmos. Pois esse grupo, em especial, apresenta as duas condições de recomendação final, o que pode vir a apresentar um limiar para aceitação de artigos que foram rejeitados no congresso.

4.2.2.1 Experimento 2-A: geração de subgrupos no grupo de desempenho heterogêneo obtido no Experimento 1-A

Aplicando-se o algoritmo x-means para o grupo DH originado no experimento 1-A, onde a base de dados utilizada foi de avaliações de artigos de todos os institutos à exceção do IFS, obteve-se a seguinte configuração para os seus subgrupos:

- DH: subdividido em três grupos, sendo eles g0, g1, g2.

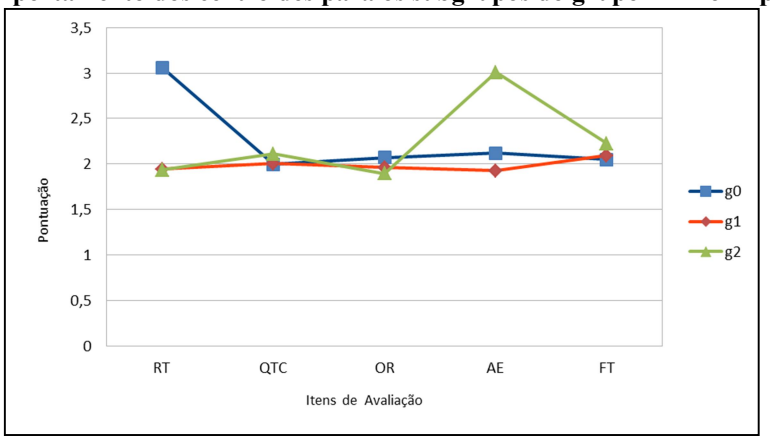
Após a aplicação do x-means no grupo DH os dados ficaram distribuídos entre 3 subgrupos, cujos centroides e quantitativo de itens encontram-se na Tabela 14.

Tabela 14 - Centroides dos subgrupos originados pelo x-means para os subgrupos do grupo DH do Experimento 1-A

Grupo	RT	QTC	OR	AE	FT	Total Itens	Total Itens (%)
g0	3,063	1,998	2,070	2,120	2,047	431	43,27
g1	1,945	2,005	1,962	1,924	2,091	419	42,07
g2	1,933	2,112	1,896	3,007	2,224	146	14,66

Fonte: Autor, 2013

Figura 9 - Comportamento dos centroides para os subgrupos do grupo DH no Experimento 1-A



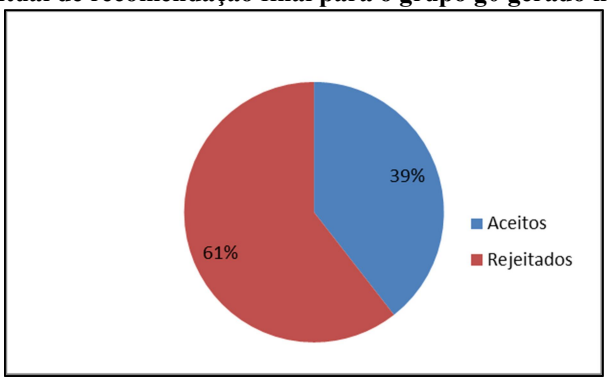
Fonte: Autor, 2013

Distribuídos de acordo com os centroides da Tabela 14, os subgrupos gerados do grupo DH possuem a tendência em manter as notas em torno de 2, comportamento esse observado no grupo que deu origem aos mesmos, sendo que após essa divisão em subgrupos duas particularidades ganham destaque. A primeira é que o subgrupo g0 foge ao comportamento citado acima para o atributo Relevância do Tema mantendo a nota em torno de 3. Outra particularidade é que mesmo acontece com g2 para o item de avaliação Apresentação e Estrutura.

Com a obtenção dos subgrupos pela mineração e observando-se o percentual de aceitos e rejeitados por subgrupo criado, notou-se que os subgrupos gerados comportam-se de maneira semelhante quanto ao aspecto de aceitação e rejeição dos artigos do grupo que os originou.

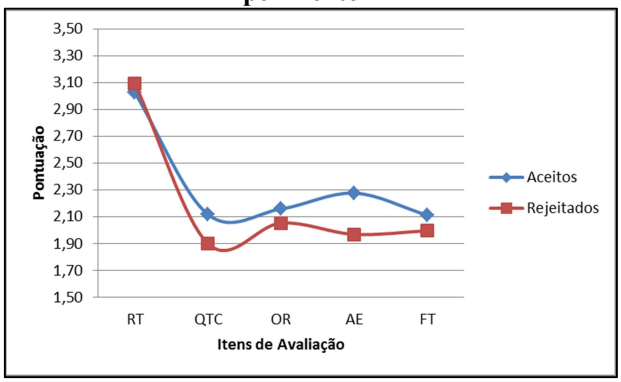
O grupo g0 conforme Gráfico 21, mantendo o comportamento do grupo DH, alocou uma maior quantidade artigos rejeitados. Das 431 avaliações que aí foram colocadas pelo algoritmo 170 são de avaliações de artigos aceitos contra 261 de avaliações de artigos rejeitados.

Gráfico 21 - Percentual de recomendação final para o grupo g0 gerado no Experimento 2-A



Fonte: Autor, 2015

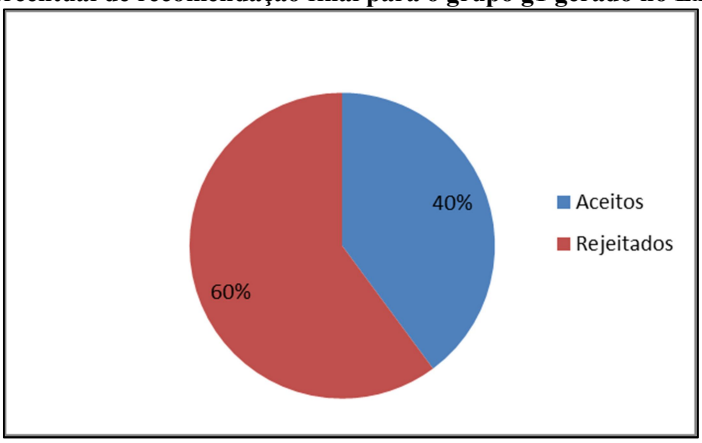
Figura 10 - Comportamento das avaliações quanto a aceitação ou rejeição no subgrupo g0 originado no Experimento 2-A



Fonte: Autor, 2015

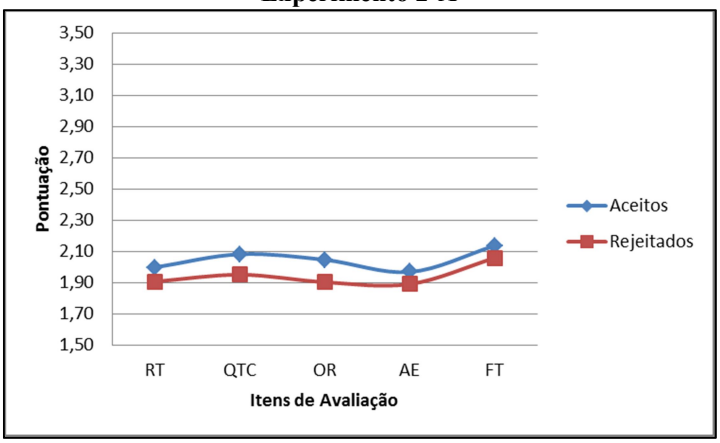
Para o grupo g1, que possui 167 (40%) de avaliações de artigos aceitos e 252 (60%) avaliações de artigos rejeitados, conforme Gráfico 22, se percebe um comportamento constante em torno da nota média 2 para todos os itens de avaliação.

Gráfico 22 - Percentual de recomendação final para o grupo g1 gerado no Experimento 2-A



Fonte: Autor, 2015

Figura 11 - Comportamento das avaliações quanto a aceitação ou rejeição no subgrupo g1 originado no Experimento 2-A



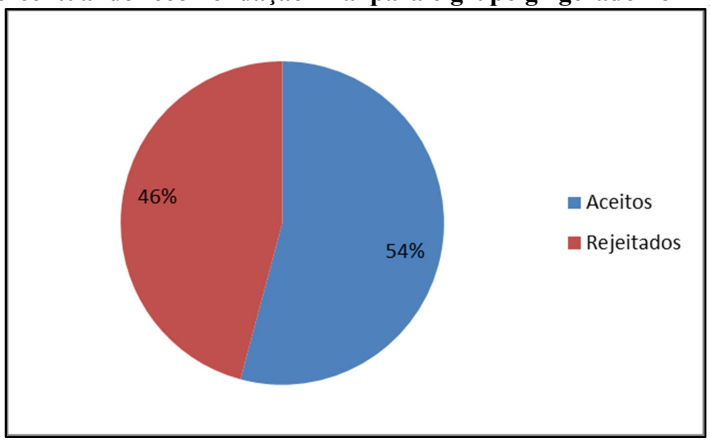
Fonte: Autor, 2015

Para o último grupo criado nesse experimento, o grupo g2, o algoritmo alocou 78 avaliações (54%) de artigos aceitos e 66 avaliações (46%) de artigos rejeitados, conforme Gráfico 23. Dos três subgrupos criados esse é o único que possui em sua maioria avaliações aceitas, divergindo do comportamento do grupo DH.

Ainda é possível verificar que o item de avaliação Apresentação e Estrutura possuiu um desvio de comportamento em relação à média dos demais, ficando assim em torno de 3. É perceptível também que nesse grupo estão alocados os melhores artigos, segundo os avaliadores e suas notas nos itens de avaliação, conforme pode ser visto na Figura 12. Aí se percebe que a média dos itens de avaliação está sensivelmente maior do que a média

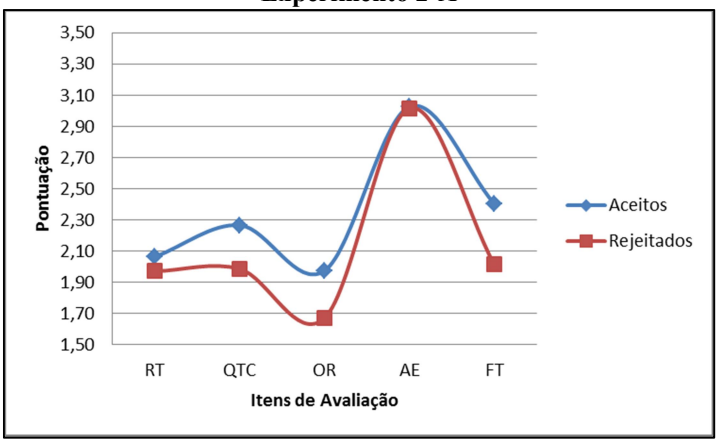
verificada nos grupos g0 e g1.

Gráfico 23 - Percentual de recomendação final para o grupo g2 gerado no Experimento 2-A



Fonte: Autor, 2015

Figura 12 - Comportamento das avaliações quanto a aceitação ou rejeição no subgrupo g2 originado no Experimento 2-A



Fonte: Autor, 2015

4.2.2.1.1 Breves conclusões a respeito dos subgrupos obtidos no Experimento 2-A

Após a aplicação do algoritmo x-means em um dos grupos resultantes do Experimento 1-A não foram possíveis grandes esclarecimentos acerca do grupo de Desempenho Heterogêneo em questão. Visto que o mesmo seguiu a mesma tendência de notas para os artigos observada no grupo os originou, o grupo DH.

Foi possível verificar uma tendência do atributo AE possuir valores levemente mais elevados dos demais no subgrupo g2, onde continha um percentual maior de aceitos. Tal comportamento não foi observado nos outros dois subgrupos gerados desse experimento.

4.2.2.2 Experimento 2-B: geração de subgrupos no grupo de desempenho heterogêneo obtido no Experimento 1-B

Ao aplicar o algoritmo x-means no grupo DH originado da base de dados de avaliações de artigos do Instituto Federal de Sergipe, obteve-se a seguinte configuração de subgrupos:

- DH: subdividido em dois subgrupos, sendo eles g0 e g1.

O grupo DH originou dois subgrupos em seu interior cujos centroides e quantitativos de itens encontram-se na Tabela 15.

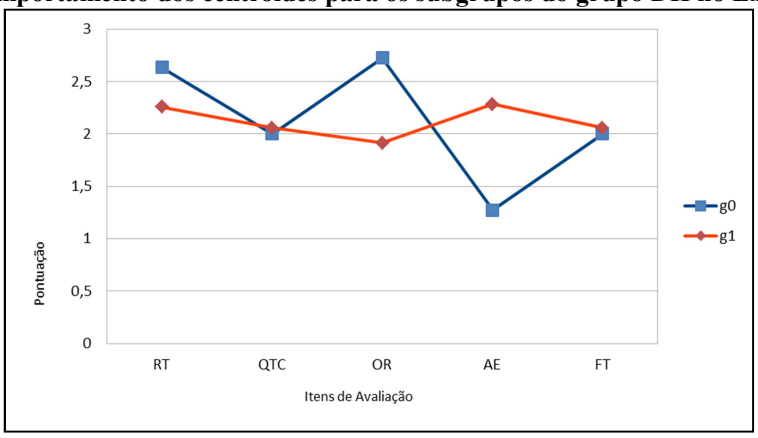
Tabela 15 - Centroides dos subgrupos originados pelo x-means para os subgrupos do grupo DH do Experimento 1-B

Grupo	RT	QTC	OR	AE	FT	Total Itens	Total Itens (%)
g0	2,636	2	2,727	1,273	2	11	23,9
g1	2,257	2,057	1,914	2,286	2,057	35	76,1

Fonte: Autor, 2015

Ambos os subgrupos originados no Experimento 2-B possuem o comportamento semelhante ao observado no Experimento 2-A. Tais subgrupos seguem a mesma tendência observada no seu grupo originário, o grupo DH do Experimento 1-B. Tendência em manter em torno de 2 a pontuação para os atributos de avaliação dos artigos.

Figura 13 - Comportamento dos centroides para os subgrupos do grupo DH no Experimento 1-B



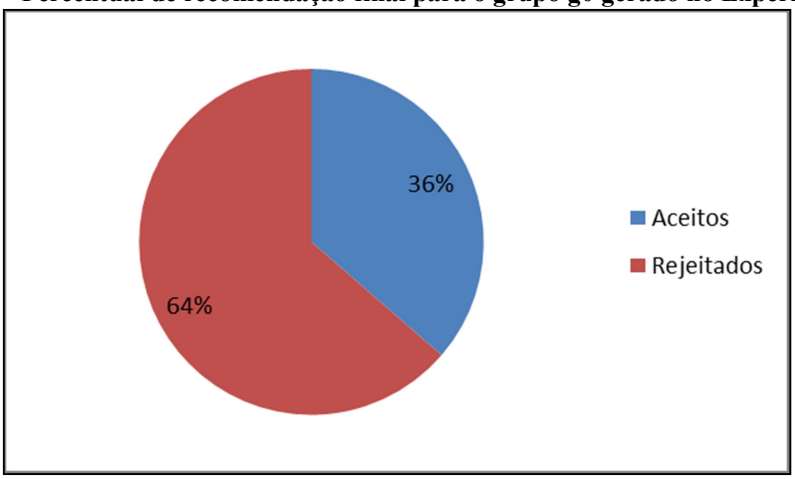
Fonte: Autor, 2015

Também com a obtenção dos subgrupos pela mineração pôde-se observar o percentual de aceitos e rejeitados por subgrupo criado, notando-se que em um subgrupo o

comportamento é o mesmo do grupo DH e em outro esse comportamento se inverte.

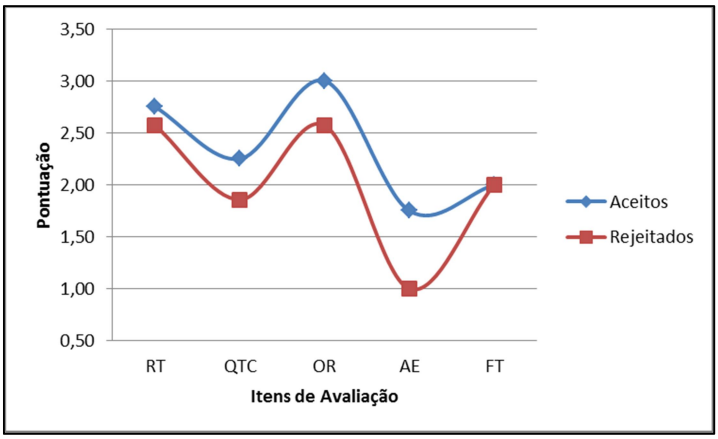
O grupo g0 de acordo com o Gráfico 24, alocou uma maior quantidade artigos rejeitados. Expressando comportamento oposto, quanto à quantidade de aceitos e rejeitados, ao observado no grupo DH. Das 11 avaliações que aí foram colocadas pelo algoritmo 4 são de avaliações de artigos aceitos (36%) contra 7 de avaliações de artigos rejeitados (67%).

Gráfico 24 - Percentual de recomendação final para o grupo g0 gerado no Experimento 2-B



Fonte: Autor, 2015

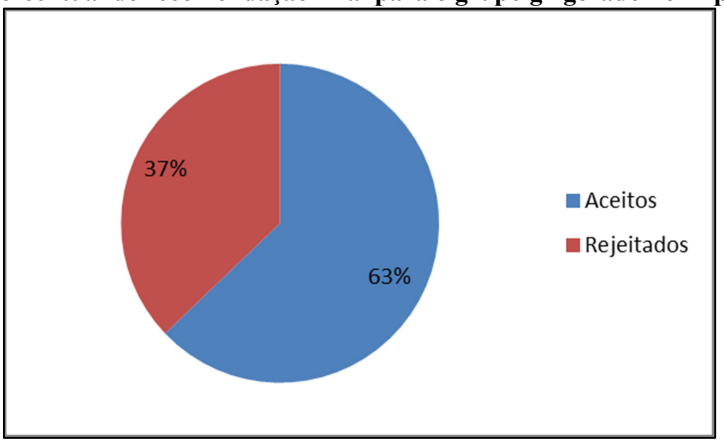
Figura 14 - Comportamento das avaliações quanto à aceitação ou rejeição no subgrupo g0 originado no Experimento 2-B



Fonte: Autor, 2015

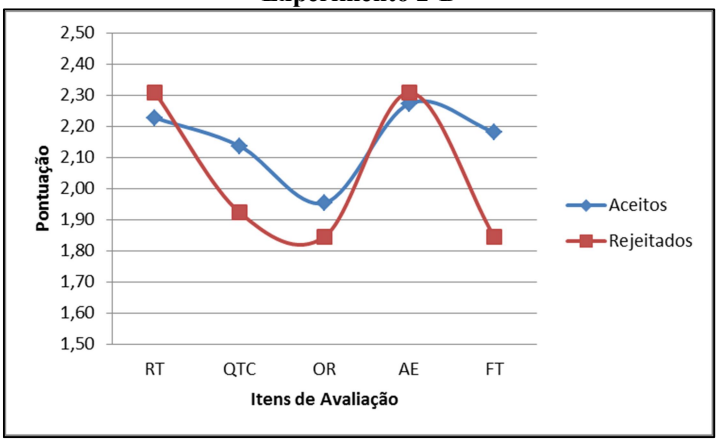
No subgrupo g1, a maioria dos itens é de avaliações de artigos aceitos 22 (63%) enquanto possui 13 avaliações de artigos rejeitados (37%), conforme visto no Gráfico 25, expressando comportamento semelhante ao grupo DH.

Gráfico 25 - Percentual de recomendação final para o grupo g1 gerado no Experimento 2-B



Fonte: Autor, 2015

Figura 15 - Comportamento das avaliações quanto à aceitação ou rejeição no subgrupo g1 originado no Experimento 2-B



Fonte: Autor, 2015

4.2.2.2.1 Breves conclusões a respeito dos subgrupos obtidos no Experimento 2-B

De maneira similar ao Experimento 2-A, a aplicação do algoritmo no grupo DH originado no Experimento 1-B não elucidou os motivos que levaram a aceitação ou rejeição dos artigos no congresso em questão. Ele trouxe subgrupos com a mesma característica do grupo que os originou, a tendência em manter em torno de 2 a pontuação para os itens de avaliação.

Observou-se também nesse experimento a tendência seguida pelo atributo AE, também notada no Experimento 2-A. Tal atributo manteve uma pontuação um pouco mais elevada frente os demais no subgrupo g1, subgrupo esse com maior percentual de artigos aceitos. Tal tendência não foi observada no subgrupo g0, subgrupo com maior percentual de artigos rejeitados.

5 CONCLUSÕES

O trabalho aqui realizado de busca por conhecimento na base de dados através da caracterização por meio de técnica de agrupamento é extensível a outras instituições que procurem caracterizar a produção de seus pesquisadores por meio de submissões em eventos. Na presente dissertação contempla-se um estudo de caso do Instituto Federal de Sergipe no Congresso Norte Nordeste de Pesquisa e Inovação.

Após os experimentos propostos e realizados no capítulo 3 e capítulo 4, e das análises[18] sobre os grupos e subgrupos gerados, é possível a partir deste capítulo identificar as características e assim fornecer informações, com base nessas características dos perfis definidos pelos agrupamentos gerados, que irão auxiliar a PROPEX e o Instituto Federal de Sergipe a alavancar e consolidar o desenvolvimento da atividade de pesquisa na instituição.

5.1 Perfis dos Trabalhos Propostos no VI CONNEPI

As propostas submetidas pelos pesquisadores ao evento foram avaliadas segundo os critérios: Relevância do tema, Qualidade Técnico-científica, Originalidade, Apresentação e Estrutura e por fim Fundamentação Teórica. Cabendo ao avaliador expressar sua recomendação final sobre o trabalho a partir de sua análise dos itens supracitados.

Portanto, procura-se nesse tópico definir características do perfil do pesquisador por instituição com base na avaliação de seus artigos por meio dos critérios, acima citados, devidamente julgados pelos avaliadores do evento. Devendo esse servir de subsídio a fim de realizar um trabalho de intervenção por parte da PROPEX personalizado por grupo gerado de pesquisadores pertencentes ao Instituto Federal de Sergipe mediante os demais institutos federais de educação, ciência e tecnologia do país participantes do congresso.

5.1.1 Características dos trabalhos propostos ao congresso pelos pesquisadores dos institutos federais de ciência e tecnologia do país à exceção do IFS

Nesse grupo de submissões ocorreu a presença marcante da resposta “Bom” para os itens avaliados, além de possuir o item relevância do tema como sendo o mais bem pontuado perante os demais itens. Esse item também possuiu a maioria do percentual de notas “Bom”

---

[18] Realizadas no capítulo 4.

frente os demais. Essa informação apenas ratifica a presença dos artigos no congresso, pois os mesmos devem estar alinhados ao tema central do evento, não fugindo assim à sua proposta.

Assim, mediante os experimentos realizados na base de dados de avaliações de artigos usada no Experimento 1-A (todos os institutos à exceção do IFS) é possível perceber algumas características na avaliação dos trabalhos submetidos ao congresso por esse grupo de pesquisadores.

Primeiramente a formação de quatro grupos distintos de acordo com o valor das notas para os itens de avaliação que os mesmos abrangem, sendo nomeados como DA, DB, DR e DH, conforme Tabela 11.

Tal formação permite claramente a separação dos artigos submetidos ao evento de acordo com a recomendação final dada pelos avaliadores de artigos do congresso. À exceção do grupo DH que tanto possui dados de submissões de artigos aceitos como rejeitados, tendo a média dos itens de avaliação em torno de 2. O fato de possuir avaliações com perfis aceito e rejeitado motivou a realização do Experimento 2-A.

Podemos observar no grupo DR a presença das notas em sua maioria valoradas como 1 pelos avaliadores do congresso, exceto o item relevância do tema destacando-se dos demais, seguido do item originalidade. Cabendo ao item qualificação técnico científica a pior avaliação nesse grupo. Assim, o grupo DR é um grupo que possui as piores avaliações do evento e que contém notas acima da média para o item relevância do tema e as piores notas do congresso para o item qualificação técnico científica.

Originado no processo de mineração do Experimento 1-A, o grupo DH, apresentou a maior heterogeneidade quanto ao aceite ou rejeição de artigos no congresso, característica essa que o fez receber esse nome. Apresentou também um comportamento semelhante ao grupo DR quanto ao atributo relevância do tema. Tal atributo destacou-se em média dos demais itens de avaliação dentro do grupo.

Quanto aos demais itens de avaliação, estes se mantiveram mais constantes do que no grupo DR, encontrando-se a média das notas para os itens desse grupo em torno de 2. Coube ao item originalidade o pior desempenho dentre os demais itens no grupo DH.

Grupo com 51,89% da quantidade de avaliações de artigos no congresso, o grupo DB alocou artigos com notas em torno de 3, mantendo um tendência constante em torno desse valor para os itens avaliados. Assim, o grupo DB apresentou a maior quantidade de avaliações alocadas pelo processo de mineração dos quatro grupos formados, tendo em média um bom

desempenho e sendo aceitos os artigos referentes às avaliações aí alocadas.

Por fim, o grupo DA é possuidor das avaliações de maiores notas no congresso. Com média em torno de 3,5 esse grupo apresentou a tendência de ser constante nas notas dos itens de avaliação, também observada nos grupos DH e DB. A única leve queda em média se dá para o item originalidade, que mesmo assim não possui grande distância dos demais itens no grupo.

Com a realização do Experimento 2-A, motivada pela existência do grupo DH, obtivemos 3 subgrupos a fim de elucidar características desse grupo que possui bem próxima, em percentual, a quantidade de artigos aceitos e rejeitados.

O subgrupo g0 que possui comportamento constante para os itens de avaliação em torno de 2, exceto para o item relevância do tema cuja média está em torno de 3. O subgrupo g1 possuidor de um comportamento constante em torno de 2 para todos os itens de avaliação. E por fim o subgrupo g2, que apresentou um comportamento contrário do grupo DH, no que diz respeito ao percentual de artigos aceitos e rejeitados, possuindo sua maior parte de artigos aceitos.

No subgrupo g2, grupo com percentual maior de artigos aceitos em relação aos rejeitados no grupo DH, existe a tendência das notas do atributo apresentação e estrutura ser mais alta perante os demais. Comportamento esse não observado em g0 e g1. Assim existe uma leve tendência que leva a crer que esse atributo pode ter colaborado na aceitação dos artigos presentes em DH.

5.1.2 Características dos trabalhos propostos ao congresso pelos pesquisadores do Instituto Federal de Sergipe

Comportando-se de maneira similar ao grupo de avaliações de artigos trabalhado no item 5.1.1, o grupo de avaliações de artigos submetidos ao congresso por pesquisadores do IFS possui em sua maioria a avaliação “Bom” para os itens, registrando essa uma frequência de 52,41%, conforme Tabela 9.

Assim como o grupo de artigos de todos institutos à exceção do IFS o atributo relevância do tema também manteve a maior nota frente os demais itens para as avaliações de artigos do IFS. Tais características mostram uma proximidade no desempenho, visto de maneira estatística, dos artigos do IFS perante os demais institutos do país.

Da mesma forma após a aplicação do algoritmo de mineração na base de dados de artigos do IFS obtivemos quatro grupos: DA, DH, DB, DR. Grupos esses que foram nomeados de acordo com o perfil de notas das avaliações que abrigavam. Novamente a nota do item relevância do tema é a maior dentre as demais em cada grupo originado no experimento.

O grupo DA, possuidor das melhores notas de avaliações para os artigos do IFS, teve para o item originalidade a pior pontuação mediante os demais itens. Manteve também a média em torno de 3,5 para as notas dos atributos de avaliação. Média essa também verificada no grupo DA de avaliação de artigos de todos os institutos.

Também presente nas avaliações dos artigos do IFS, o grupo DB, possui a característica percebida anteriormente, um equilíbrio no valor de suas notas para os atributos e também médias um pouco menores em relação ao grupo DA.

Nos grupos DH e DR dos artigos do IFS observa-se também o comportamento notado para as avaliações dos artigos dos demais institutos, notas estáveis próximas à média dos grupos.

Assim como o grupo DR das avaliações dos artigos à exceção do IFS, o grupo DR de avaliações do IFS apresentou um comportamento diferente para o atributo relevância do tema, que superou a média de nota para os demais dentro do grupo. Nesse caso os artigos do IFS com desempenho ruim no congresso possuíram por vezes notas maiores para esse item do que os artigos que estão alocados no grupo DH. E coube ao atributo qualidade técnico científica a menor pontuação para os artigos com notas mais baixas do IFS no congresso.

Também motivada pela presença do grupo DH, foi realizado o experimento 2 para os artigos do IFS. Assim, obtivemos os subgrupos g0 e g1, diferentemente o ocorrido para os artigos dos demais institutos que foi gerado do experimento 2 a quantidade de três subgrupos.

O subgrupo g0 obtido demonstra comportamento contrário quanto à quantidade de aceitos e rejeitados, ao observado no grupo DH, possuindo em seu interior maioria de artigos rejeitados. Nesse subgrupo observa-se também o comportamento do item apresentação e estrutura sendo o possuidor das notas mais baixas nesse subgrupo. Fato que leva a uma leve tendência dos artigos que aí estão terem sido rejeitados por conta desse atributo.

Já no subgrupo g1, que apresenta maioria de artigos aceitos em seu interior seguindo inclusive a tendência do grupo DH, percebe-se a permanência em torno da média para os itens de avaliação dos artigos, tendo o item apresentação e estrutura a tendência em acompanhar

essa média.

5.2 Ações Sugeridas para o Desenvolvimento da Pesquisa no Instituto Federal de Sergipe Baseadas nas Informações Obtidas no Presente Trabalho

Mediante as características das submissões dos artigos de pesquisadores do IFS no VI CONNEPI trazidas por esse trabalho e a comparação com as submissões realizadas pelos demais institutos federais do país no mesmo congresso é possível fornecer subsídio através das informações aqui trabalhadas para que ações que auxiliem a gestão no desenvolvimento da pesquisa no Instituto Federal de Sergipe sejam propostas pela PROPEX. Buscando alcançar assim bons resultados frente o cenário nacional, equilibrando assim a produção acadêmica desenvolvida nesta instituição e dando retorno ao incentivo proposto pelo governo federal, para que dessa forma ainda mais pesquisas possam ser desenvolvidas na instituição através da assistência e fomento prestados pelo mesmo e por outras agências financiadoras.

Assim, a PROPEX possui um conhecimento das características desses grupos e pode atuar em ações personalizadas por grupo gerado no processo de mineração.

Conforme descrito e observado, é possível subsidiar a PROPEX em uma ação de melhoria da apresentação e estrutura dos artigos submetidos ao evento, que possuíram um comportamento peculiar frente os demais itens no grupo de desempenho heterogêneo mediante o que foi levantado no Experimento 2-B. Sendo proposto então um aperfeiçoamento na forma minicurso afim de aprimorar e orientar os pesquisadores da instituição no desenvolvimento metodológico de seus artigos em publicação, reforçando dessa maneira conceitos que os auxiliem a propor suas ideias de maneira organizada e apresentá-las ao público nos mais diversos eventos pelo país.

Uma segunda ação embasada por esse trabalho e que serve de sugestão a ser tomada diz respeito a um maior incentivo e apoio à participação em eventos pelo país mesmo sem a publicação do pesquisador no evento, através de uma maior aplicação da verba destinada a pesquisa a custear viagens e participações em congressos. Dessa forma o pesquisador estaria em constante contato com o que vem sendo desenvolvido em termos de sua área de atuação e isso o auxiliará no tocante ao aprimoramento da originalidade de seus trabalhos, visto que esse ponto apresentou-se com um desempenho muito inconstante nos mais diversos grupos detectados pelo processo de descoberta de conhecimento realizado no presente trabalho.

Tal sugestão também aprimora o desenvolvimento do pesquisador no tocante a fundamentação teórica de seus trabalhos, item de baixa pontuação na avaliação dos artigos do

grupo DH. Pois dessa maneira o pesquisador teria contato constante com as novas pesquisas que estão sendo desenvolvidas em sua área de atuação, o que o auxiliaria a embasar sua pesquisa na instituição.

5.3 Conclusão e Trabalhos Futuros

Após a utilização do processo de descoberta de conhecimento foi possível, no presente trabalho, identificar características que antes não eram percebidas apenas com um trabalho estatístico aplicado sobre os dados do VI CONNEPI. Pôde-se então identificar com mais precisão o perfil dos trabalhos dos pesquisadores do Instituto Federal de Sergipe submetidos ao evento e assim compará-lo ao perfil desenvolvido pelos demais institutos federais de educação ciência e tecnologia do país, observando dessa maneira o equilíbrio existente no IFS frente o cenário nacional.

Alguns outros trabalhos podem ser realizados em a fim de fornecer novas ações de subsídio para desenvolvimento da atividade de pesquisa no Instituto Federal de Sergipe. Dentre eles podem se destacar:

- Fazer uso da área de atuação do pesquisador para verificar em quais áreas há uma fragilidade dos artigos submetidos nos itens julgados pelos avaliadores no congresso, acrescentando-a assim ao perfil descrito no item 5.1.2, o que possibilitará agir diretamente nos cursos com fragilidades no desenvolvimento da pesquisa;
- Utilizar os projetos de pesquisa submetidos à PROPEX nos anos de 2012, 2013 e 2014 para uma análise evolutiva, até então sem interferência da pró-reitoria, do desenvolvimento e amadurecimento das ideias trabalhadas em pesquisa na instituição pós VI CONNEPI. Assim, realizar a aplicação das ações citadas no item 5.3 e verificar a evolução na avaliação nos artigos submetidos em eventos e os próprios projetos submetidos à PROPEX no ano de 2015.

REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data mining**. Addison Wesley Longman, England, 1996.

ANDRADE, Maria Margarida de. **Introdução à metodologia do trabalho científico**: elaboração de trabalhos na graduação. 5. ed. São Paulo: Atlas, 2001.

BERRY, M. J. A.; LINNOFF, G. **Data Mining techniques** – for marketing, sales, and customer support. 3. Ed. United States: Wiley Computer Publishing, 2011.

BRASIL. Ministério da Educação e Cultura. CONAES. **Diretrizes para a avaliação das instituições de educação superior**. Brasília: INEP, 2004.

BRASIL. Ministério da Educação e Cultura. **GEOCAPES**. Disponível em: <http://geocapes.capes.gov.br/geocapesds/#>. Acesso em: 12 jun. 2013.

BRASIL. Lei n. 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica, cria os Institutos Federais de Educação, Ciência e Tecnologia, e dá outras providências. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 30 dez. 2008. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm>. Acesso em: 04 ago. 2015.

BRASIL, Thâmara; GUEDES, Silvio; PAGANINE, Joseana. TEIXEIRA, J. Carlos. **Produção científica avança.** Revista em Discussão, Brasília, p. 26-28 set. 2012. Disponível em: <http://www.senado.gov.br/noticias/jornal/emdiscussao/Upload/201203%20-%20setembro/pdf/em%20discuss%C3%A3o!_setembro_2012_internet.pdf> Acesso em: 28 ago. 2013.

CARVALHO, L. A. V. de. **Data mining: a mineração de dados no marketing, medicina, economia, engenharia e Administração**. São Paulo: Érica, 2001.

DIAS, M. A. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados**. 2001. 212 f. Tese85 (Doutorado do Programa de Pós-Graduação em Engenharia de Produção) - UFSC. Florianópolis, Santa Catarina, 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery**: an overview. In: Advances in knowledge Discovery and data mining, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, 1996.

GOLDSCHMIDT, R.; PASSOS, E. **Mineração de Dados**: um guia prático. Rio de Janeiro: Campus, 2005.

HAN, Jiawei; KAMBER, Micheline. **Data Mining Concepts and Techniques.** 2. Ed. São Francisco: Elsevier, 2006.

INSTITUTO FEDERAL DE SERGIPE. **Regulamento de Pesquisa e Extensão do Instituto Federal de Sergipe.** Aracaju. 14 ago. 2011. Disponível em: <http://www.ifs.edu.br/images/divcom/2011/Arquivos/Setembro/regulamento%20de%20pesquisa%20e%20extenso.pdf >. Acesso em: 20 set. 2013.

INSTITUTO FEDERAL DE SERGIPE. **Relatório de Autoavaliação Institucional da Comissão Própria de Avaliação do IFS.** Aracaju. 2012. Disponível em: < http://www.ifs.edu.br/sistemas?id=1186>. Acesso em: 20 set. 2013.

INSTITUTO FEDERAL DE SERGIPE. **Relatório de Gestão do Exercício 2012 da Pró-reitoria de Pesquisa e Extensão do IFS.** Aracaju. 2013.

KUMAR, Vipin; STEINBACH, Michael; TAN, Pang. **Introdução ao datamining**. Rio de Janeiro: Ciência Moderna LTDA, 2009.

MARTINS, Gilberto de Andrade. **Estatística geral e aplicada**. 3. Ed. São Paulo: Atlas, 2009.

MITCHELL, Tom M. **Machine learning**. McGraw-Hill, 1997.

OLIVEIRA, Adelize G. de. **Data Warehouse: Conceitos e Soluções**. Florianópolis: Editora Advanced, 1998. 96p.

PELLEG, Dan; MOORE, Andrew W. **X-means**: Extending K-means with Efficient Estimation of the Number of Clusters. In: Seventeenth International Conference on Machine Learning, 727-734, 2000.

QUINLAN, J. R**. Learning efficient classification procedures and their application to chess end games.** In J.G. Carbonell, R. S. Michalski, and T.M. Mitchell, editors: Machine Learning, v.1, Tioga, Palo Alto, USA, 1983.

RIGHETTI, Sabine. **Produção científica no Brasil aumenta, mas cai qualidade.** Folha de São Paulo, São Paulo, 22 abr. 2013. Disponível em: <http://www1.folha.uol.com.br/fsp/cienciasaude/105099-producao-cientifica-do-brasil-aumenta-mas-qualidade-cai.shtml#_=_ > Acesso em: 28 ago. 2013.

RUSSEL, S.; NORVIG, P.; **Inteligência Artificial**. 2 Ed. São Paulo: Editora Campus, 2004.

SILVA, Ermes Medeiros da, et al. **Estatística para os cursos de administração, ciências contábeis e economia**. 2. Ed. São Paulo: Atlas, 1996.

SPIEGEL, Murray R. **Probabilidade e Estatística**. São Paulo: Pearson Education do Brasil, 1978.

UNIVERSIDADE FEDERAL DE SERGIPE. **Anuário estatístico da Universidade Federal de Sergipe.** São Cristóvão. 2012. Disponível em: <http://oficiais.ufs.br/sites/default/files/21/anuario_estatistico_da_ufs_2011.pdf>. Acesso em: 04 ago. 2015.

APÊNDICE

APÊNDICE A - X-means implementado na ferramenta RapidMiner

O RapidMiner possui em seu núcleo a implementação do algoritmo de agrupamento x-means desenvolvido por Dan Pelleg e Andrew Moore e publicado no ano 2000. A implementação realizada utiliza-se do BIC (Bayesian Information Criteria) para determinar a quantidade grupos a ser dividida a base de dados. Além de conter também os demais benefícios do algoritmo de acordo com a publicação "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" (MOORE; PELLEG, 2000).

Dentre as características do algoritmo implementado no RapidMiner, pode-se citar:

Entrada

- **example set**: *expects:* ExampleSetMetaData: #examples: = 0; #attributes: 0 , *expects:* ExampleSet

Saída

- **cluster model**
- **clustered set**

Parâmetros

- k min: O menor número de grupos que devem ser divididos os dados.
- k max: O maior número de grupos que devem ser divididos os dados.
- measure types: o tipo de medição usado para alocar dados nos grupos.