UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JEAN CARLO DE BORBA ESPÍNDOLA

Um Estudo Analítico Sobre o Comportamento de Pesquisadores Baseado em Dados de Produção Científica

Trabalho de Graduação.

Prof. Dr. Renata Galante
Orientador

Prof. Msc. Cristiano Roberto Cervi
Co-orientador

Porto Alegre, dezembro de 2009.

AGRADECIMENTOS

Agradeço à minha família pela educação, exemplo de vida e por tudo o que me ensinaram na vida. Aos meus orientadores pelas ajudas inestimáveis. Á minha namorada Maiara pela compreensão, carinho e incentivo.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DCBD	Descoberta de Conhecimento em Base de Dados
HTML	HyperText Markup Language
IDAMS	Internationally Developed Data Analysis and Management Software
KDD	Knowledge Discovery Database
UNESCO	United Nations Educational, Scientific and Cultural Organization
WEKA	Waikato Environment for Knowledge Analysis
XML	eXtensible Markup Language
XSLT	eXtensible Stylesheet Language Transformations

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

As informações sobre a vida científica de pesquisadores tem se mostrado um importante instrumento para a análise e tomada de decisões envolvendo produções científicas. Neste contexto, a Plataforma Lattes apresenta-se como uma importante base de dados consolidada, destinada ao fomento da pesquisa científica e formação de recursos humanos para pesquisa científica no país.

O objetivo deste trabalho é aplicar técnicas de mineração de dados para analisar e avaliar a produção científica de um grupo de pesquisadores da área da ciência da computação através das informações existentes no currículo Lattes. A abordagem se dá através de análise estatística dos dados obtidos do Lattes, bem como pela utilização de técnicas de análise de dados, como agrupamento e regressão, para mostrar o comportamento de pesquisadores através de artigos publicados, orientações de mestrado e doutorado, rede de co-autoria e seu nível de produção.

As principais contribuições do estudo são a análise temporal do comportamento científico de pesquisadores brasileiros da área de banco de dados com a aplicação de técnicas de análise de dados sobre a produção destes pesquisadores, identificando o comportamento dinâmico da produção nos períodos analisados.

Palavras-Chave: Lattes, mineração de dados, análise de produção científica.

ABSTRACT

The information about the scientific life of researchers has been an important tool for the analysis and decision making involving scientific production. In this context, the Lattes Platform is important as a consolidated database, for the promotion of scientific research and training of human resources for scientific research in the country.

The objective of this work is to apply techniques of data mining to analyze and evaluate the scientific work of a group of researchers in the field of computer science through the information obtained in Lattes. The approach is through statistical analysis of data from the Lattes and the use of techniques of data analysis such as clustering and regression to show the behavior of researchers through published articles, master's and doctoral's guidelines, co authorship network and its level of production.

The main contributions of the study are: the temporal analysis of the scientific behavior of Brazilian database area researchers with the application of data analyzing techniques on the production of these researchers, identifying the dynamic behavior of production in the periods that have been analyzed.

Keywords: Lattes, data mining, analysis of scientific production.

1 INTRODUÇÃO

Atualmente a gestão da ciência, da tecnologia e da inovação passa por um processo de qualificação, uma vez que se busca cada vez mais conhecer os pesquisadores e suas atividades, com vistas à avaliação qualitativa de suas pesquisas financiadas por diversos órgãos de fomento. Esta avaliação serve, também, como orientação para uma correta aplicação dos recursos de fomento à pesquisa.

A tarefa de avaliar a produção científica de um pesquisador é baseada fortemente na análise dos dados disponíveis em seu currículo. Por outro lado, a plataforma Lattes[1], apesar de ser o mais importante instrumento de armazenamento dos dados referentes à produção científica de pesquisadores brasileiros não possui um mecanismo que permita consultas mais elaboradas sobre os mesmos. Dessa forma, todo o processo de análise é realizado de forma humana, demandando muito tempo, sendo extremamente cansativo e podendo levar a avaliações equivocadas, pois toda análise fica baseada somente na percepção humana de um volume muito grande de dados. Esta deficiência é devido à falta de ferramentas que ofereçam a possibilidade de se realizarem consultas diretas de recuperação de informações que permitam análise e cruzamento de dados. Um dos fatores que tornam esta tarefa mais difícil é que a produção científica de um pesquisador é construída ao longo do tempo, sendo que uma análise agregando todos os dados oferece uma visão distorcida da realidade.

A análise temporal do currículo Lattes de pesquisadores é fundamental para o processo de tomada de decisão. Agências como Capes[2] (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e CNPq[3] (Conselho Nacional de Desenvolvimento Científico e Tecnológico), que necessitando avaliar a produção científica de pesquisadores brasileiros, precisam de mecanismos semi-automatizados para apoiar as avaliações e oferecer uma visão quantitativa da evolução do trabalho acadêmico do pesquisador ao longo do tempo. Assim, é importante a existência de ferramentas de apoio para analisar a rede de co-autoria de pesquisadores, o nível de integração entre eles, a colaboração científica na produção conjunta, bem como orientações de mestrado e de doutorado que podem ter em comum.

A utilização da base de dados do Lattes para análise de dados de pesquisadores não é algo inédito em computação. Trabalhos recentes, como os de Arruda et al. (2009) e

[1] Plataforma Lattes. Disponível em http://lattes.cnpq.br

[2] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Disponível em http://www.capes.gov.br

[3] Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em http://www.cnpq.br

Wainer et al. (2009), utilizaram o Lattes como base de dados para fazer experimentos e realizar cruzamento de dados. O presente trabalho visa abordar questões temporais que envolvem a produção científica de pesquisadores brasileiros, esta é sua principal contribuição. A abordagem se dá através de análise estatística dos dados obtidos do Lattes, bem como pela utilização de técnicas de análise de dados, como agrupamento e regressão, para mostrar o comportamento de pesquisadores através de artigos publicados, orientações de mestrado e doutorado, rede de co-autoria e seu nível de produção.

Para a definição do grupo de pesquisadores analisados, foram identificados doutores em informática de diversas áreas de atuação, vinculados a programas de pós-graduação. Os pesquisadores foram classificados em 3 grupos, utilizando-se como critério o tempo de doutorado, ou seja, os que concluíram o doutorado até o ano de 1994, entre 1995 e 2001, e os que concluíram entre 2002 e 2008. As principais contribuições do estudo são a análise temporal do comportamento científico de pesquisadores brasileiros da área de banco de dados com a aplicação de técnicas de análise de dados sobre a produção destes pesquisadores, identificando o comportamento dinâmico da produção nos períodos analisados.

A seção 2 apresenta a base conceitual utilizada ao longo do trabalho, bem como as ferramentas utilizadas no processo de mineração de dados.. A abordagem proposta, com a arquitetura do trabalho, metodologia, critérios de classificação, forma de aquisição, extração e mineração dos dados é detalhada na seção 3. A seção 4 apresenta a análise de resultados sob o ponto de vista da análise temporal, experimentos utilizando técnicas de agrupamento e experimentos utilizando técnicas de regressão. Finalmente, na seção 5, são realizadas as considerações finais e esboçados trabalhos futuros.

2 BASE CONCEITUAL

O objetivo deste capítulo é apresentar os principais conceitos e ferramentas utilizadas durante o desenvolvimento do trabalho. A parte conceitual é composta basicamente do processo de descoberta de conhecimento em bases de dados (DCBD, ou KDD – Knowledge Discovery Database), bem como de uma análise da etapa de mineração de dados, onde são apresentados alguns métodos e algoritmos para a manipulação dessa etapa. Também são apresentadas as ferramentas que se aproveitam destes conceitos para realizar a mineração de dados.

2.1 Processo de KDD e a Mineração de Dados

A descoberta do conhecimento em base de dados é a obtenção de conhecimento e informações úteis a partir de grandes conjuntos de dados armazenados. O processo de KDD, bem como a mineração de dados, é caracterizado por várias etapas, que se tornam necessárias para a obtenção do resultado desejado, conforme apresentado na Figura 2.1.

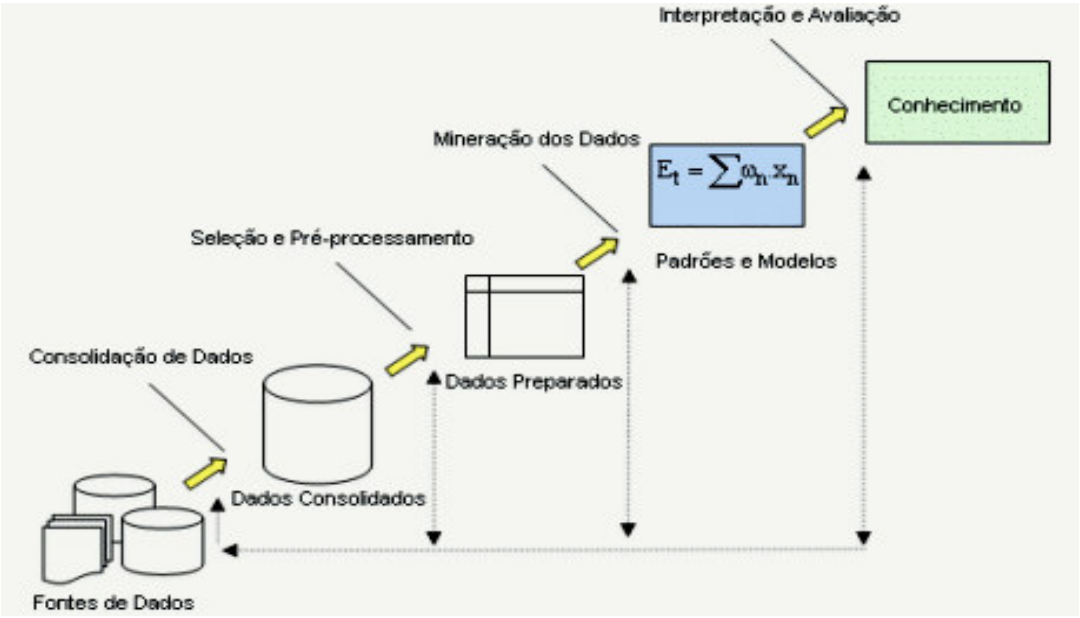

Figura 2.1: Etapas do processo de KDD (FAYYAD et al., 1996).

As etapas do processo de KDD são: Definição dos Objetivos, Seleção de Dados, Pré-Processamento, Transformação, Mineração de Dados e Interpretação e Avaliação. Conforme Fayyad (1996) as etapas podem ser resumidas em: Seleção, Pré-processamento, Transformação, Mineração de Dados e Interpretação/Validação. Desse esquema, autores como Brusso (2000) e Goldschmidt e Passos (2005), concordam em agrupar ainda mais as etapas, tornando o processo de KDD dividido em apenas três

etapas operacionais, sendo estas: Pré-Processamento, Mineração de Dados e Pós-Processamento.

2.1.1 Pré-processamento

A etapa de pré-processamento inclui operações básicas de manipulação de dados, tais como remover ruídos ou subcamadas, coletando informação necessária para as tarefas de descoberta de conhecimento. Além disso, nessa etapa são decididas estratégias para manusear (tratar) campos que não são necessários no processo posterior de mineração de dados.

Uma descrição mais detalhada dessas operações abrange os seguintes itens (GOLDSCHIMDT e PASSOS, 2005):

- Seleção de Dados: função que compreende a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD;
- Limpeza de Dados: envolve qualquer tipo de tratamento referente aos dados selecionados de forma a assegurar a qualidade da informação por eles representada;
- Codificação dos Dados: os dados devem ser codificados de forma que possam ser utilizados como entrada para os algoritmos de mineração de dados. Por exemplo, transformar valores numéricos em categorias ou intervalos; ou o inverso, como representar numericamente valores de atributos categóricos;
- Transformação e Enriquecimento dos Dados: A transformação consiste em reduzir as variáveis envolvidas por atributos que representem as características principais dos dados das bases. Já o enriquecimento é um item onde se buscam mais informações que possam ser agregadas a base de dados, para prover melhores resultados. Para isso, podem-se realizar consultas a bases externas de dados, pesquisas de campo e outras técnicas (GOLDSCHIMDT e PASSOS, 2005).

2.1.2 Mineração de Dados

A mineração de dados consiste em uma etapa de KDD responsável pela seleção dos métodos a serem utilizados para detectar padrões nos dados (FAYYAD et al., 1996). Parte desse processo está na busca por padrões de interesse em uma forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para realizar uma tarefa (SILVA, 2004).

As tarefas de mineração de dados podem ser classificadas em duas categorias, de acordo com Han e Kamber (2006): padrões descritivos e padrões preditivos. As tarefas descritivas caracterizam as propriedades gerais dos dados em uma base de dados. Já os padrões preditivos buscam através de inferências predizer valores desconhecidos. Padrões preditivos são construídos para resolver problemas específicos de predição, em um ou mais atributos de um banco de dados. Ainda afirma que esses padrões não precisam necessariamente prever o futuro, mas sim fazer a descoberta de um atributo desconhecido a partir de outros atributos conhecidos. Em contrapartida, os padrões descritivos não têm por objetivo resolver um problema específico, mas apresentar padrões interessantes que o especialista em um domínio desconhece (JOHN, 1997). Two Crows Corporation (2005) apresenta a definição de que padrões descritivos são

geralmente usados para criar subgrupos significativos, como agrupamentos de dados demográficos, por exemplo.

2.1.2.1 Métodos Preditivos

- Classificação: função que mapeia (classifica) um registro de uma base de dados a uma ou mais classes (rótulos) pré-definidos. Feito isso, é possível aplicar essa função a novos registros que venham ser incorporados à base de dados. As formas mais comuns de representação de conhecimento dos algoritmos de classificação são regras e árvores.

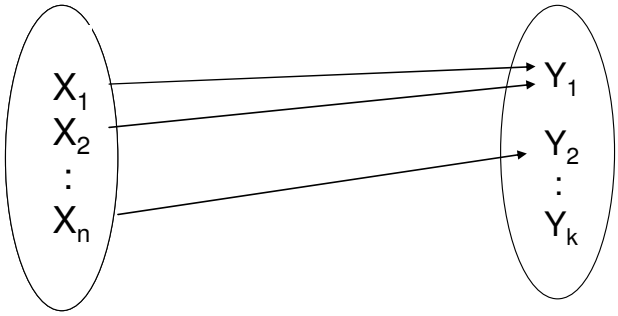

Figura 2.2: Mapeamento entre registros e classes de uma base de dados (GOLDSCHIMDT e PASSOS, 2005).

Na Figura 2.2, é representado um exemplo de mapeamento entre registros e classes de maneira que dois registros X1 e X2 podem ser mapeados para um registro Y1.

Um exemplo de classificação é o caso de uma base de dados com os seguintes atributos: nome, sexo, total de atuações profissionais e número de orientações feitas por pesquisadores, como na Tabela 2.1.

Tabela 2.1: Exemplo de conjunto de dados para aplicação de Classificação

	NOME	**SEXO(M/F)**	**ATUAÇÕES**	**ORIENTAÇÕES**
1	A	M	6	9
2	B	M	5	32
3	C	F	5	50
4	D	M	3	60
5	E	M	2	40
6	F	M	3	9
7	G	M	2	22
8	H	M	3	77
9	I	M	2	0
10	J	M	6	33
11	K	M	1	63
12	L	M	2	27
13	M	F	10	38
14	N	M	1	54

A partir dos dados expressos na Tabela 2.1, aplicando o algoritmo de classificação J48[4], a saída obtida na forma de árvore de decisão na Figura 2.3.

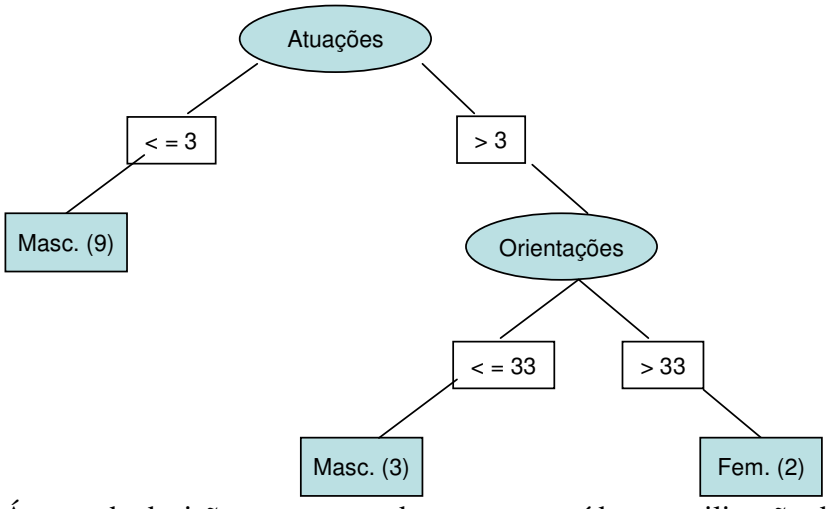

Figura 2.3: Árvore de decisão representando regras extraídas na utilização do algoritmo J48.

Descrevendo-se de formar textual a informação contida na árvore de decisão da Figura 2.3, pode-se inferir que: (i) quando o total de atuações profissionais de uma pessoa é menor ou igual a três esta pessoa é do sexo masculino; e (ii) se as atuações forem maiores que três e o número de orientações feitas por pesquisadores for maior ou igual a trinta e três, as pessoas pertencem ao sexo masculino, caso contrário se maiores que trinta e três, as pessoas pertencem ao sexo feminino.

- Regressão: consiste na busca por uma função que mapeia os registros de um banco de dados em valores reais (GOLDSCHIMDT e PASSOS, 2005). Essas funções podem ser lineares ou não, e se restringem a dados numéricos. Um exemplo dessa técnica é o mapeamento de informações de tempo de experiência (variável X) em anos e salário anual dos funcionários de uma empresa (variável Y). Aplicando-se o modelo de regressão linear pode-se chegar a uma função linear. Dessa maneira, baseado nos dados armazenados, obtêm-se uma função linear que pode ser utilizada para predizer o valor de uma variável em função da outra. Por exemplo, pode-se formular a seguinte questão, cuja resposta pode ser prevista por meio da aplicação de uma função obtida com dados anteriores:
    - Quantos anos de experiência é preciso para ganhar X salário anual, com base nos valores passados que a empresa armazena?

2.1.2.2 Métodos Descritivos

- Agrupamento: é a classificação de objetos ou registros em diferentes grupos, ou seja, particionamento do conjunto de dados em subconjuntos ou agrupamentos, assim cada subconjunto compartilha similaridades. Diferente da tarefa de classificação que possui rótulos pré-definidos, no agrupamento é o algoritmo que identifica automaticamente os rótulos (FAYYAD et al., 1996). Os itens de um conjunto de dados são agrupados de acordo com alguma métrica de

4 J48: Implementação em Java do algoritmo de árvore de decisão C4.5 encontrada na ferramenta Weka. Disponível em: <www.cs.waikato.ac.nz/ml/weka/>.

similaridade[5], como a distância euclidiana. Assim, padrões pertencentes a um agrupamento devem ser mais similares entre si do que em relação aos outros agrupamentos. É função do usuário definir a quantidade de agrupamentos a serem representados. A análise que pode ser feita após essa definição é quais são as características comuns aos elementos de cada agrupamento, a fim de definir um rótulo que o represente. Na Figura 2.4 tem-se uma representação de agrupamentos definidos através do algoritmo K-Means (MACQUEEN, 1967; HUANG, 1998), que trabalha somente com dados numéricos e não categóricos:

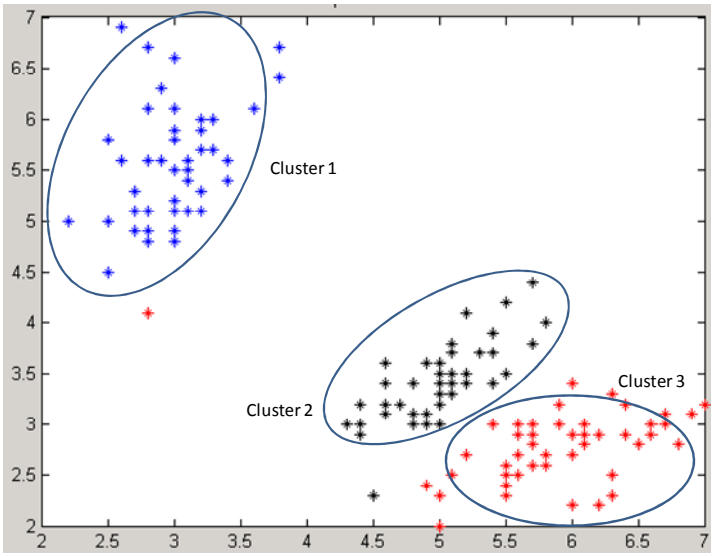

Figura 2.4: Representação visual de três agrupamentos gerados pelo algoritmo K-Means (MATHWORKS, 2008).

- Sumarização: envolve métodos para encontrar uma descrição das características comuns para conjuntos de dados ou agrupamentos obtidos através da técnica de agrupamento. O uso de técnicas de sumarização ajuda na compreensão de um domínio (NATARAJAN e SHEKAR, 2005). Isso pelo fato de a sumarização levantar características essenciais de um domínio. Entretanto, de maneira geral, ela não pretende buscar nada novo. Isso ocorre porque as descrições reveladas fazem parte de um conhecimento já difundido no domínio do usuário, tornando-se uma técnica que permite apenas uma melhor visualização da informação, deixando de lado a descoberta por novos padrões e interesses que estão implícitos nos dados;
- Regras de Associação: esta tarefa tem por objetivo encontrar padrões freqüentes entre conjunto de itens de dados e é muito usada no âmbito das aplicações comerciais e na área de marketing. Isso porque através do uso dessas regras é possível obter, por exemplo, a análise da freqüência com que um cliente adquire um item A e também leva um item B na mesma compra. Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Estudo sobre extração de regras de associação na mineração da web foi desenvolvido por Brusso

[5] Métrica de similaridade: determina o quão parecidos são dois conjuntos de caracteres. As métricas são avaliadas por um valor que varia de 0 a 1 ([0;1]) onde 0 implica que são totalmente diferentes e 1 que são idênticos. Esse valor é calculado com base em uma métrica de distância a ser escolhida dependendo da função de similaridade utilizada.

(2000), visando obter os padrões de comportamento de usuários ao navegarem em páginas web. Dois algoritmos que implementam a tarefa de descoberta de regras de associação são: Apriori (AGRAWAL e SRIKANT, 1994) e C4.5 (QUINLAN, 1993).

2.1.3 Pós-Processamento

Após o pré-processamento e a mineração de dados, a etapa de pós-processamento é de suma importância para que o resultado final seja obtido.

Esta é a fase que engloba a visualização, a análise e a interpretação das saídas geradas na etapa de Mineração de Dados. Com isso, é gerado um modelo de conhecimento, ou seja, qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva algum conjunto de dados (GOLDSCHIMDT e PASSOS, 2005; FAYYAD et al. 1996).

2.2 WEKA

WEKA[6] (*Waikato Environment for Knowledge Analysis*) é uma ferramenta de KDD, construída na Universidade de Waikato na Nova Zelândia, que compreende diversos algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. Nela há também mecanismos para filtragem e normalização de dados, através dos chamados "filters". Conforme detalhamento feito por Eibe et al. (2005), a ferramenta Weka possui algoritmos que podem ser agrupados nas seguintes categorias:

- Classificação: nessa categoria estão implementados os mais usados e conhecidos algoritmos de classificação. Existem métodos bayesianos, árvores de decisão como ID3 (QUINLAN, 1979 e GOLDSCHMIDT; PASSOS, 2005), C4.5 (QUINLAN, 1993) e aprendizagem de regras como OneR;
- Regressão: estão implementados diversos algoritmos de regressão, tais como regressão linear simples e múltipla, pace regression, regressão vetorial, perceptron multicamadas, locally-weighted learning e árvores de decisão;
- Clusterização: apenas alguns algoritmos estão implementados, tais como, K-Means (MACQUEEN, 1967 e HUANG, 1998), algoritmos EM (Estimation-Maximization), Cobweb e Farthest-First Clustering;
- Regras de associação: estão presentes os algoritmos Apriori, PredictiveApriori e Tertius;
- Seleção de atributos: são suportadas as abordagens filter e wrapper (GOLDSCHIMDT; PASSOS, 2005):
    - Filter: Trata-se de uma abordagem onde a seleção de atributos independe do algoritmo de mineração que será aplicado aos atributos selecionados;
    - Wrapper: Consiste em experimentar o algoritmo de mineração de dados para cada conjunto de atributos e posteriormente avaliar os resultados obtidos.

[6] Weka. Disponível em: www.cs.waikato.ac.nz/ml/weka/

As implementações de seleção de atributos na Weka, incluem seleção baseada em correlação, estatística qui-quadrado, gain ratio, information gain, symmetric uncertainty e vector machine-based criterion. Existem também uma variedade de métodos de pesquisa, como seleção seqüencial para trás (backward selection) e seleção seqüencial para frente (forward selection), algoritmo de busca best-fit, genetic search algorithm e pesquisa aleatória.

- Filters: processos que transformam instâncias e conjuntos de instâncias são chamados de "filters", e são classificados de acordo com o que faz mais sentido, apenas em um contexto de predição (chamado supervisionado) ou em qualquer contexto (chamado não-supervisionado). Além disso, são divididos em "attribute filters", que trabalham em um ou mais atributos de uma instância, e "instance filters", que trabalham em todas as instâncias. Conforme Eibe et al. (2005, p. 6), as tarefas desempenhadas em cada um desses filtros são:
    - Filtros de atributos não-supervisionados incluem adicionar um novo atributo, adicionar um indicador de cluster, adicionar ruído, copiar um atributo, discretizar um atributo numérico, normalizar e padronizar um atributo numérico, mesclar valores de atributos, transformar valores nominais em binários, ofuscar valores, trocar valores, remover atributos, substituir valores nulos [...]

Os filtros de atributos supervisionados incluem suporte à seleção de atributos, discretização[7], transformação nominal para binária, reordenação de valores de classe. Na questão dos filtros de instâncias não-supervisionadas, estes transformam instâncias esparsas em não-esparsas e vice-versa. Também realizam a tarefa de randomizar e fazem resampling[8] em conjuntos de instâncias, além de remover instâncias de acordo com certos critérios (EIBE et al., 2005, p. 6-7).

A ferramenta Weka suporta a manipulação de diferentes formatos de arquivos como CSV, binário e C45, bem como, possibilita a busca direta de dados de servidores de banco de dados e também de servidores Web. O formato nativo da Weka é o ARFF (Attribute-Relation File Format), que é um arquivo ASCII usado para definir atributos e seus dados (WEKA, 2009). A apresentação dos resultados é disponibilizada em forma gráfica de histogramas, árvores de decisão, diagramas de dispersão, além de prover modelos gráficos para montagem de redes neurais (GOLDSCHIMDT; PASSOS, 2005).

2.3 IDAMS

O IDAMS[9] (***I**nternationally **D**eveloped **D**ata **A**nalysis and **M**anagement **S**oftware*) é um pacote de software para validação, manipulação e análise estatística de dados. É

[7] Discretização de dados: consiste em transformar atributos contínuos, que representam valores reais, de uma base de dados em atributos discretos, que podem assumir apenas valores pertencentes a um determinado conjunto. (VOLTOLINI, 2006).

[8] Resampling: Métodos estatísticos os quais são capazes de estimar a taxa de erro verdadeira de forma confiável mesmo para pequenas amostras de casos resolvidos. (VOLTOLINI, 2006).

[9] IDAMS. Disponível em: <http://portal.unesco.org/ci/en/ev.php-URL_ID=2070&URL_DO=DO_TOPIC&URL_SECTION=201.html>.

organizado como uma coleção de facilidades de gerência de dados e análise acessíveis através de uma interface de usuário e uma linguagem de controle comum.

Distribuído sem custos pelo Secretariado da UNESCO[10] (*United Nations Educational, Scientific and Cultural Organization*), também pode ser obtido através dos distribuidores oficiais é originalmente derivado do pacote de software OSIRIS III.2 desenvolvido no Instituto para Pesquisa Social da Universidade de Michigan, U.S.A. Foi e continua sendo enriquecido continuamente pelo secretariado da UNESCO com a cooperação de especialistas de diferentes países, como Estados Unidos, Bélgica, Colômbia, França, Hungria, Polônia, Rússia, Eslováquia e Ucrânia.

Constitui em uma ferramenta poderosa para profissionais que, em ambiente científico, educacional ou administrativo, têm que manusear e analisar dados numéricos.

Utilizando conceitos de KDD, oferece grande variedade de técnicas de análise de dados, tais como, técnicas de regressão (regressão linear simples e regressão linear múltipla), análise de variância, técnicas de *clustering,* em uma interface de usuário gráfica moderna, permite a personalização do ambiente de aplicação através de um editor para criação de arquivos com instruções para a execução do programa. Oferece também um visualizador de resultados com informações pertinentes aos cálculos dos algoritmos executados.

[10] UNESCO. Disponível em: < http://portal.unesco.org>.

3 MÉTODO PROPOSTO

Este capítulo descreve o método proposto que tem como objetivo analisar a produção científica de pesquisadores da área de ciência da computação. Para isso, foi usada uma abordagem temporal, utilizando como fonte de informação o histórico da produção dos pesquisadores. Os indicadores empregados na verificação de produção foram o número de co-autores, o número de publicações em conferências, o número de publicações em periódicos e a quantidade de orientações em nível de mestrado e de doutorado de cada pesquisador. O intervalo de produção analisado foi um período de 9 anos, compreendendo os anos de 2000 a 2008.

Este capítulo está organizado da seguinte forma. Na primeira seção é apresentada a arquitetura da implementação desenvolvida. Em seguida, é descrita a metodologia utilizada para o desenvolvimento do trabalho e o processo de extração de dados. Por fim, são detalhadas as técnicas de mineração de dados utilizadas para analisar a evolução dos pesquisadores.

3.1 Arquitetura

Nesta seção, é descrita a arquitetura geral utilizada para implementação da proposta descrita neste trabalho. A figura 3.1 ilustra a arquitetura proposta que é constituída de 4 grandes etapas: metodologia, extração dos dados, mineração dos dados e análise dos resultados.

Na etapa “Metodologia”, foram definidos aspectos iniciais do trabalho como tamanho e integrantes da base de dados, a subdivisão feita para os mesmos, baseado no tempo de pesquisa de cada um dos integrantes e o período de produção a ser analisado para os pesquisadores.

Em “Extração de Dados”, foram definidas as formas que as informações necessárias seriam extraídas da Web e posteriormente preparadas para a próxima etapa do trabalho.

Na etapa “Mineração de Dados”, foram definidos os métodos de análise e mineração que os dados passariam para, então, terem os resultados analisados. Os métodos escolhidos foram: análise temporal utilizando como ponto de comparação indicadores dos pesquisadores; algoritmos de agrupamento utilizando os indicadores como variável de comparação; e algoritmos de regressão, utilizando indicadores dos pesquisadores pra predizer resultados de produção.

Por fim, a etapa de “Análise de Resultados” compreende a análise e comparação de todos os resultados do processo de Mineração de Dados, buscando a identificação de padrões e comportamento dos pesquisadores e dos grupos definidos.

As próximas seções descrevem detalhadamente cada uma das etapas da arquitetura proposta.

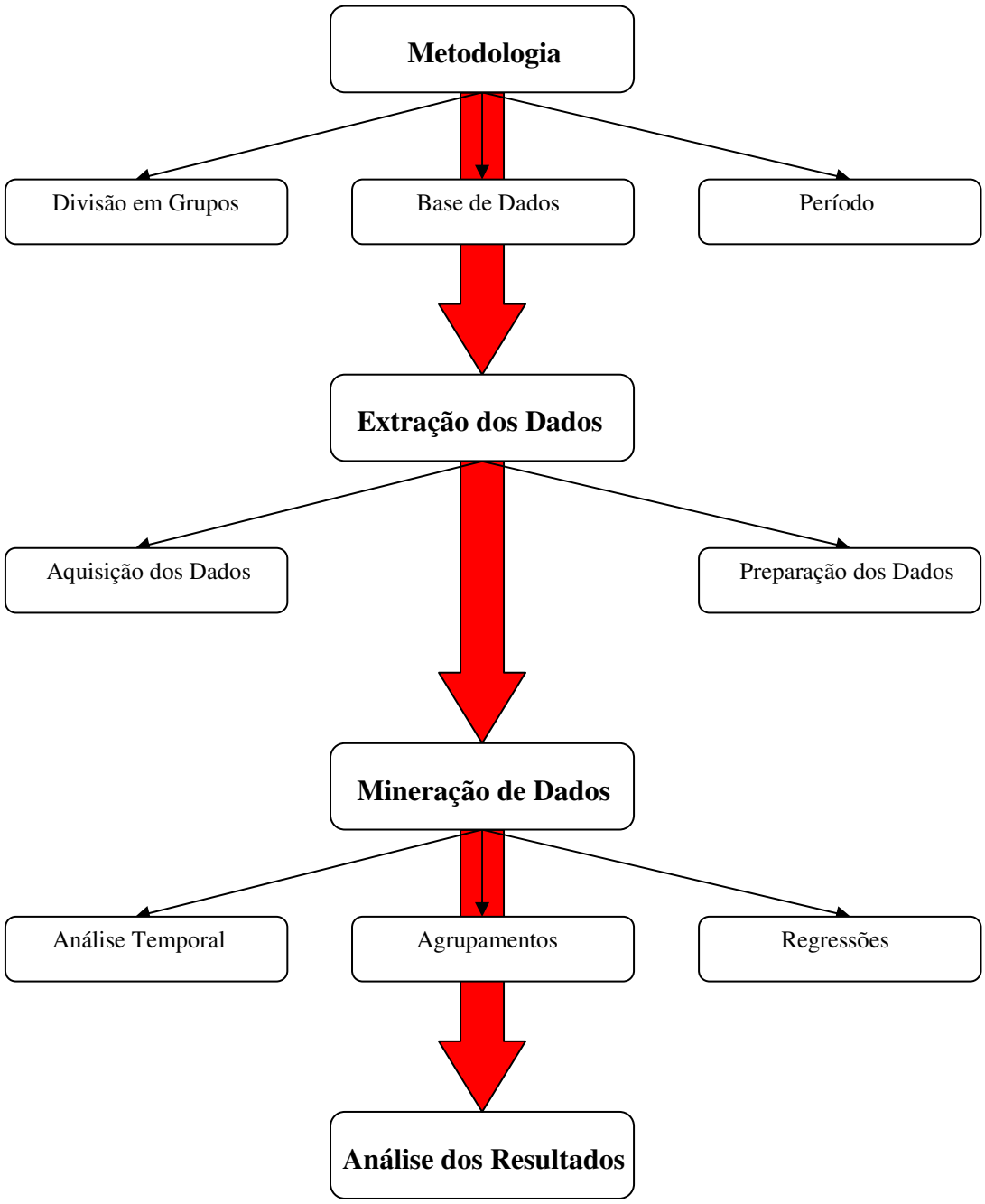

Figura 3.1: Arquitetura geral do trabalho.

3.2 Metodologia

Esta seção descreve as definições iniciais do trabalho indispensáveis para o desenvolvimento do mesmo. São explicados como foram selecionados os pesquisadores para comporem a base de dados, o processo de subdivisão dos grupos e a seleção dos períodos analisados.

3.2.1 Base de Dados

Foram selecionados 228 pesquisadores doutores brasileiros da área de Ciência da Computação, que concluíram o doutorado até o ano de 2008. Estes pesquisadores são oriundos de diversas instituições nacionais, todos vinculados a programas de pós-graduação *Stricto Sensu*, com conceito Capes 5, 6 e 7. Os pesquisadores foram selecionados tendo como critério sua atuação, quer seja como orientadores em programas de pós-graduação *Stricto Sensu* reconhecidos pela Capes, como membros de comitês de programas de conferências e ainda como autores e co-autores na área.

3.2.2 Definição dos Grupos

Os pesquisadores selecionados na seção 3.2.1 foram divididos em 3 grupos, assim distribuídos:

- Grupo 1 (G1): 121 pesquisadores que concluíram o doutorado antes de 1994 (inclusive). Esses pesquisadores são identificados em uma sequência de P1 a P121 e foram classificados com elevada experiência em pesquisa;
- Grupo 2 (G2): 81 pesquisadores que concluíram o doutorado entre 1995 (inclusive) e 2001 (inclusive). Estes pesquisadores são identificados em uma sequência de P122 a P202 e foram classificados com significativa experiência em pesquisa;
- Grupo 3 (G3): 26 pesquisadores que concluíram o doutorado entre 2002 (inclusive) e 2008 (inclusive). Estes pesquisadores são identificados em uma sequência de P203 a P228 e foram classificados com razoável experiência em pesquisa.

O resultado da divisão desigual entre os grupos mostra que os pesquisadores foram escolhidos de forma a contemplar o tempo de titulação máxima, ou seja, o ano em que concluíram o doutorado.

3.2.3 Período Analisado

A análise dos dados da produção científica dos pesquisadores foi realizada tomando-se como base 3 períodos de tempo assim definidos:

- Tempo 1 (T1): compreende o período de 2000 a 2002;
- Tempo 2 (T2): compreende o período de 2003 a 2005;
- Tempo 3 (T3): compreende o período de 2006 a 2008.

3.3 Extração dos Dados

Esta seção descreve como os dados dos pesquisadores foram extraídos e preparados para passarem pelo processo de mineração de dados.

3.3.1 Aquisição de Dados

Os dados utilizados para a análise da produção científica foram oriundos do Currículo Lattes de cada um dos 228 pesquisadores. Para isso foi utilizada uma ferramenta que realizou os processos de extração, tratamento e validação dos dados. Cada currículo foi analisado utilizando-se o padrão público de visualização

disponibilizado no sítio do CNPq no formato HTML (HyperText Markup Language). A ferramenta recebe como entrada o link público para o currículo de cada pesquisador e percorre as páginas de cada pesquisador gerando como resultado um arquivo no formato XML (eXtensible Markup Language). Esse arquivo contém todos os dados do currículo de cada pesquisador. A conversão de formato permite a realização consultas e cruzamento de dados diretamente sobre o arquivo XML, bem como a utilização dos dados por algoritmos e ferramentas de mineração de dados, facilitando a manipulação.

A partir dos 228 currículos no formato XML, foi utilizada outra função da mesma ferramenta de extração e tratamento dos dados. A função consiste em processar o arquivo XML gerado usando XSLT gerando um novo arquivo XML para cada um dos currículos já extraídos, que, de forma a facilitar a consulta aos indicadores de cada pesquisador, exibindo de forma sucinta, apenas as informações de número de co-autores, trabalhos publicados, artigos publicados, dissertações orientadas e em andamento, teses orientadas e em andamento, para os 3 períodos de tempo definidos em 3.2.3.

3.3.2 Preparação dos Dados

Após a aquisição de dados, os mesmos foram preparados para passarem pela metodologia de análise dos resultados. Para isso, foram gerados 3 arquivos base: (i) um para a ferramenta Weka; (ii) um para a ferramenta Idams ; e (iii) uma planilha para a análise da evolução temporal.

Os arquivos base para as ferramentas Weka e Idams foram gerados pela mesma ferramenta de extração e tratamento dos dados, utilizado para a aquisição dos currículos. Os arquivos consistem das informações dos pesquisadores simplificadas, tais como, código, grupo, nome, totais de co-autores, totais de orientações e de produção. Cada ferramenta exigiu um arquivo diferente, devido à necessidade de formatos diferentes para cada um dos programas.

A planilha para análise da evolução temporal foi feita através da coleta manual das informações de cada um dos currículos processados. Na planilha, foram preenchidos os valores de co-autores, artigos publicados em conferências, artigos publicados em periódicos, dissertações orientadas concluídas e em andamento, teses orientadas concluídas e em andamento. Para cada um dos pesquisadores foi respeitado o valor dos indicadores para cada um dos períodos definidos.

3.4 Mineração de Dados

Para a análise da produção científica dos pesquisadores envolvidos foi empregada uma técnica de análise temporal e duas técnicas de mineração de dados. Dentre as disponíveis, as adotadas nos experimentos deste trabalho foram as de agrupamento, regressão linear (simples e múltipla).

3.4.1 Análise Temporal

A análise temporal foi utilizada para verificar se os pesquisadores evoluíram ou não ao longo dos períodos de tempo estabelecidos na seção 3.2.3. Como o período analisado foi de 9 anos, observações importantes podem ser verificadas junto ao grupo de pesquisadores selecionados.

Através da planilha obtida na seção 3.3.2, foi realizada para cada um dos indicadores (co-autores, artigos publicados em conferências, artigos publicados em eventos, dissertações orientadas concluídas e em andamento, teses orientadas concluídas e em andamento), uma análise evolutiva, para cada um dos pesquisadores contidos na base de dados. Esta análise visou encontrar padrões individuais e coletivos dentro dos grupos distribuídos, de evolução, diminuição e pequenas reduções ao longo do tempo, bem como comparações dos resultados entre os grupos, verificando a diferença de resultados entre os pesquisadores mais e menos experientes.

3.4.2 Agrupamentos

A técnica de agrupamento visou identificar os pesquisadores envolvidos em 3 grupos pelo tempo de conclusão de doutorado de cada um destes pesquisadores. Essa análise é importante para que seja possível a verificação se o tempo de conclusão de doutorado de um pesquisador tem influência significativa ou não no resultado de sua produção.

Para a realização dos experimentos com técnicas de agrupamento, foi utilizado algoritmos de mineração de dados para agrupamento incorporados na ferramenta Weka. A ferramenta utiliza conceitos de KDD e compreende diversos algoritmos de preparação de dados e validação de resultados, gerando os resultados em um visualizador de gráficos, podendo-se alternar a escolha de atributo fixo. Para o experimento, foi utilizado o algoritmo Simple K-Means usando-se a configuração de 3 clusters com atributo código fixado, ou seja, os agrupamentos foram feitos em função de co-autores, orientações e produção.

3.4.3 Regressão

As técnicas de regressão linear simples e múltipla foram utilizadas para que uma análise entre variáveis, como número de co-autores e número de orientações de mestrado e doutorado, pudessem ser identificadas como fundamentais no número da produção dos pesquisadores. Tendo-se o número de co-autores de um determinado pesquisador, bem como o número de trabalhos que ele orientou, os algoritmos de regressão podem indicar se estas variáveis impactam significativamente ou não nas publicações em conferências ou em periódicos desse pesquisador. Neste trabalho foi utilizada a variável número de co-autores como variável independente para regressão linear simples e as variáveis número de co-autores e orientações de mestrado e doutorado como variável independente para regressão linear múltipla.

Para a realização dos experimentos, foi utilizada a ferramenta para validação, manipulação e análise estatística de dados denominada Idams. Esta ferramenta também compreende diversos algoritmos de mineração de dados, como algoritmos de regressão, bem como utiliza-se de conceitos de KDD. Foi utilizado o algoritmo RegressN para os dois experimentos de regressão: regressão linear múltipla com produção como variável dependente e co-autores e orientações como variáveis independentes e; regressão linear simples com produção como variável dependente e co-autores como variável independente. O objetivo é verificar o quanto um indicador, separadamente ou em conjunto, é fundamental para predizer a produção de um pesquisador.

4 ANÁLISE DE RESULTADOS

Este capítulo descreve a análise de resultados correspondente ao método proposto que foi descrito no capítulo anterior. A avaliação experimental e análise de resultados têm por objetivo identificar, dentre os grupos definidos, a evolução da produção dos pesquisadores, bem como a diferença de características apresentadas entre os grupos. Para isso foram analisados dados de publicações em eventos e periódicos, orientações de mestrado e de doutorado e a rede de co-autoria dos pesquisadores.

4.1 Evolução Temporal

Nesta seção será analisado a evolução temporal dos pesquisadores levando-se em conta os dados dos seguintes indicadores: número de co-autores, número de artigos publicados em eventos, número de artigos publicados em periódicos, número de dissertações orientadas concluídas e em andamento e, número de teses orientadas concluídas e em andamento. Pretende-se identificar padrões de comportamento individuais e entre os grupos, observando as diferenças apresentadas nos resultados por pesquisadores com tempos distintos de participação em programas de pesquisa.

4.1.1 Número de Co-autores

Esta seção faz uma análise dos dados de cada pesquisador levando-se em consideração o seu número de co-autores. Pretende-se identificar o comportamento com relação a sua rede de colaboração científica e de integração com outros pesquisadores. As Figuras 4.1, 4.2, 4.3, 4.4 e 4.5 apresentam os gráficos resultantes da análise, no qual se identifica o número de co-autores (em uma escala de 0 a 120) dos 228 pesquisadores (P1 ao P228) em 3 períodos de tempo (T1, T2 e T3).

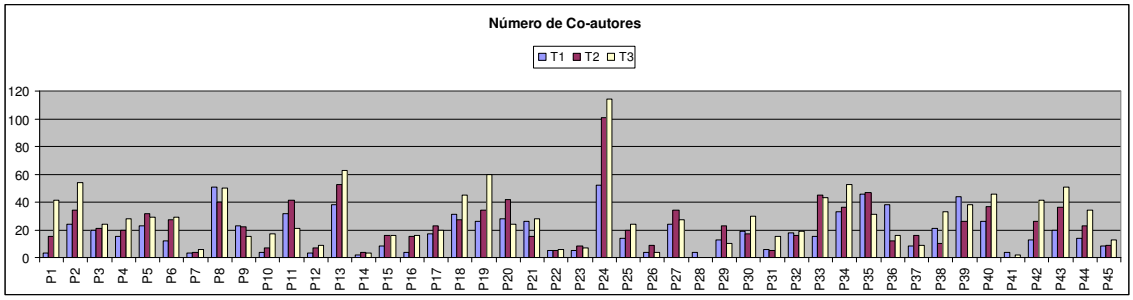

Figura 4.1: Número de co-autores de P1 à P45.

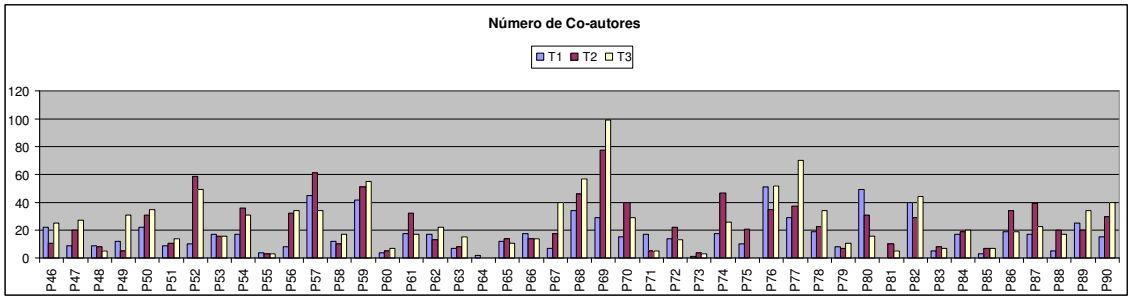

Figura 4.2: Número de co-autores de P46 à P90.

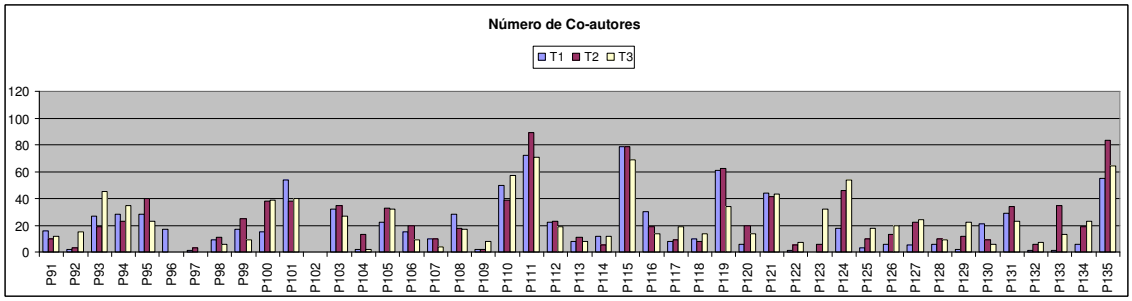

Figura 4.3: Número de co-autores de P91 à P135.

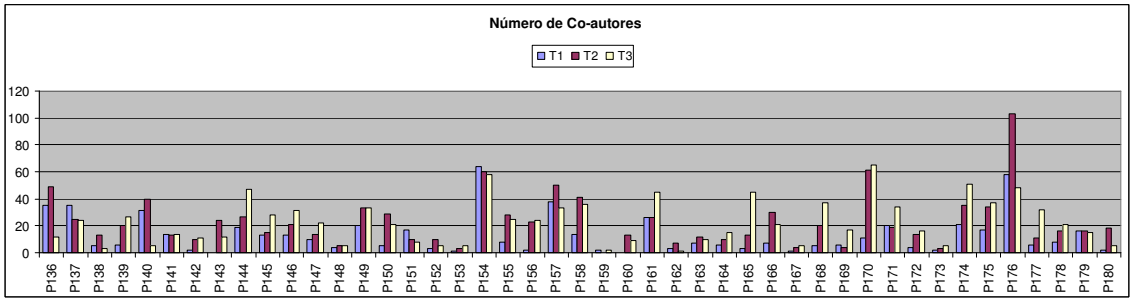

Figura 4.4: Número de co-autores de P136 à P180.

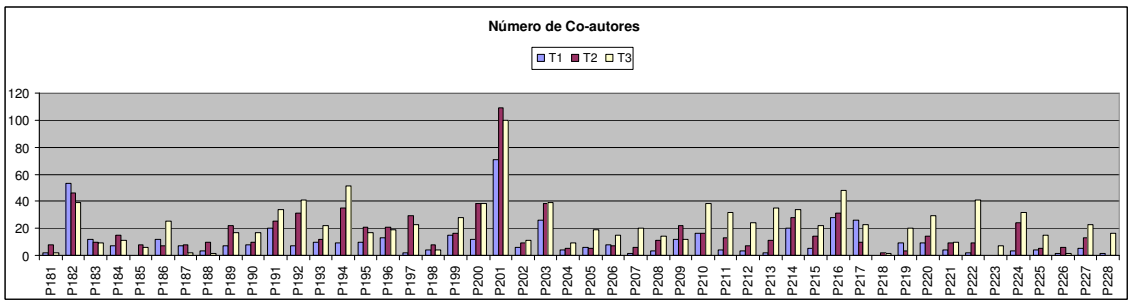

Figura 4.5: Número de co-autores de P181 à P228.

Fazendo-se uma análise dos resultados, pode-se identificar que os pesquisadores do G1 apresentaram uma rede de co-autores destacada significativamente quando comparados com os pesquisadores do G2 e principalmente do G3. Isso deve-se principalmente a sua elevada experiência em pesquisa. Entretanto, alguns pesquisadores do G2 e G3 possuem mais co-autores que alguns pesquisadores G1, tal fato se explica tanto por uma destacada posição de alguns membros do G2 e do G3, quanto uma fraca consolidação de alguns integrantes do G1. Uma explicação para o fato de os pesquisadores do grupo G1 terem obtido este resultado destacado é a consolidação das

cooperações entre os pesquisadores seniores quando as publicações concentram-se nos intercâmbios já avaliados e bem sucedidos.

Analisando-se o número de co-autores na passagem seqüencial pelos 3 períodos de tempo, 70 pesquisadores do G1, 60 do G2 e 23 do G3 apresentaram evolução. Em oposição a esta situação, 30 pesquisadores do G1, 13 do G2 e 1 do G3 diminuíram seu número de co-autores ao longo dos períodos analisados. Se analisarmos períodos não seqüenciais, 8 pesquisadores do G1 apresentaram uma redução no número de co-autores em determinados períodos de tempo.

A média de co-autores dos pesquisadores dos 3 grupos apresentou evolução, conforme especificação a seguir:

- média do G1 foi 19.31 em T1, 23.77 em T2 e 25.57 em T3;
- média do G2 foi 12.77 em T1, 22.72 em T2 e 23.01 em T3;
- média do G3 foi 7.77 em T1, 11.88 em T2 e 22.27 em T3.

Na análise realizada para co-autoria, identificou-se que, na média, os pesquisadores apresentaram um crescimento em todos os períodos de tempo analisados. Isto evidencia que os pesquisadores estão aumentando sua rede de colaboração científica, interagindo com novos colaboradores. Adicionalmente, observa-se que para G1 e G2 o maior crescimento foi identificado entre os períodos T1e T2. Isso pode ser atribuído a uma queda no número de co-autores no período T3 ou a uma manutenção da posição adquirida nos períodos anteriores.

Com relação ao G3, este grupo apresentou um crescimento maior para o número de co-autores no período T3, tendo inclusive, alguns pesquisadores aumentado seu número de co-autores de forma exponencial, ultrapassando pesquisadores com mais experiência, principalmente alguns do G1. Isto pode ser explicado pelo fato de muitos pesquisadores do G3 estarem ainda em curso no seu doutorado nos períodos de tempo T1 e T2, estando plenamente ativos apenas no período T3 onde mostraram relativo destaque, consolidando sua posição com um grande esforço de interação, e considerável crescimento perante a si mesmos.

4.1.2 Número de Artigos Publicados em Eventos

Esta seção faz uma análise dos dados de cada pesquisador levando em consideração o número de artigos publicados em eventos. Pretende-se identificar se os pesquisadores evoluíram com relação a sua produção. As Figuras 4.6, 4.7, 4.8, 4.9 e 4.10 apresentam os resultados da análise, no qual se identifica o número de artigos publicados em eventos (em uma escala de 0 a 120) dos 228 pesquisadores (P1 ao P228) em 3 períodos de tempo (T1, T2 e T3).

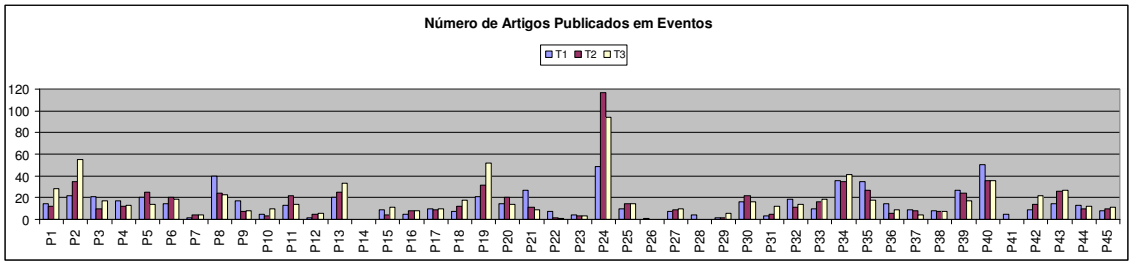

Figura 4.6: Número de artigos publicados em eventos de P1 à P45.

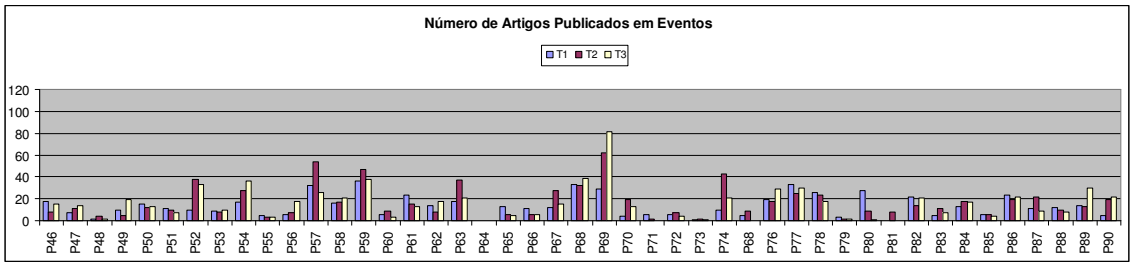

Figura 4.7: Número de artigos publicados em eventos de P46 à P90.

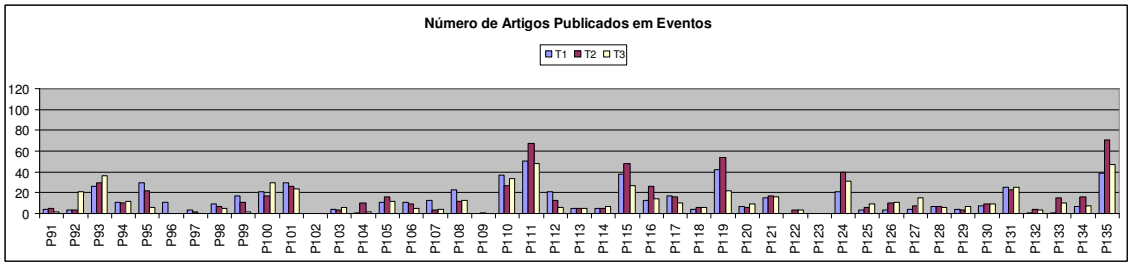

Figura 4.8: Número de artigos publicados em eventos de P91 à P135.

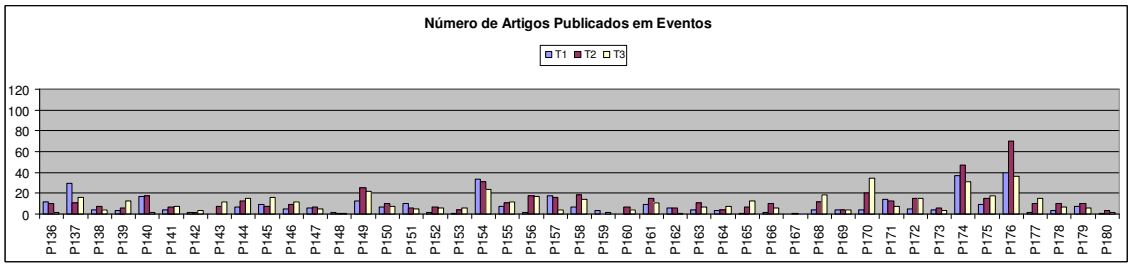

Figura 4.9: Número de artigos publicados em eventos de P136 à P180.

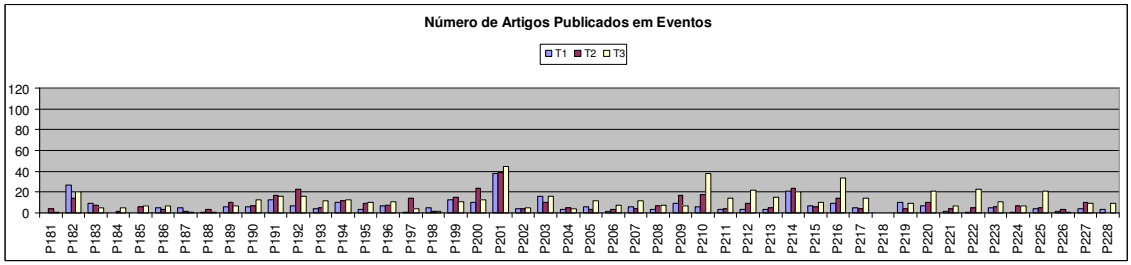

Figura 4.10: Número de artigos publicados em eventos de P181 à P228.

Os pesquisadores do G1, em sua maioria, confirmaram sua senioridade apresentando uma produção mais alta que a do G2 e do G3, quando comparados entre os três períodos de tempo. Analisando, agora, o número de trabalhos de alguns pesquisadores do G3, pode-se perceber que alguns aumentaram sua produção, no tempo T3, com relação a grande maioria dos pesquisadores do G2 e a alguns pesquisadores do G1.

Com relação à evolução nos 3 intervalos temporais, 49 pesquisadores do G1, 52 pesquisadores do G2 e 19 pesquisadores do G3 apresentaram aumento de sua produção. Em oposição a esta situação, nos 3 períodos de tempo, 51 pesquisadores do G1, 16 do G2 e 2 do G3 diminuíram sua publicação em eventos. Se analisarmos períodos não seqüenciais, 4 pesquisadores do G1 e 1 do G3 apresentaram redução quantitativa de produção em determinados períodos de tempo e nenhum pesquisador do G2 apresentou

redução quantitativa de produção na passagem pelos 3 períodos de tempo. Observa-se que está sendo analisada a produção quantitativa dos pesquisadores. Dessa forma, uma possível explicação para as reduções identificadas pode ser o fato de alguns pesquisadores estarem submetendo trabalhos para veículos mais qualificados, em conferências de abrangência internacional.

A média de trabalhos publicados em eventos dos pesquisadores dos 3 grupos apresentou os dados conforme especificação a seguir:

- média do G1 foi 14.48 em T1, 15.99 em T2 e 15.36 em T3;
- média do G2 foi 8.19 em T1, 12.31 em T2 e 10.95 em T3;
- média do G3 foi 5.42 em T1, 7.19 em T2 e 13.54 em T3.

Na análise realizada para trabalhos publicados em eventos, identificou-se que, na média, os pesquisadores apresentaram evolução positiva em todos os intervalos analisados apenas para o G3. O aumento de produção deste grupo pode ter relação com o aumento no número de co-autores dos pesquisadores, bem como com a participação efetiva dos pesquisadores, tendo em vista que nos períodos anteriores muitos deles ainda estavam com seus doutorados em curso.

Com relação ao grupo G1 e G2, ambos apresentaram uma queda de produção no período T3. No grupo G1, isso pode ser explicado pelo fato de termos mais pesquisadores que apresentaram diminuição de produção ao longo do tempo do que pesquisadores que apresentaram evolução de sua produção. Como neste grupo encontra-se uma maior senioridade, observamos também que o crescimento da média do grupo para o período T2 e decréscimo para o período T3 não foram tão acentuados, pois são pesquisadores já consolidados. Já para o grupo G2, observa-se que tivemos mais pesquisadores com evolução do que com decréscimo na produção, no entanto, esta evolução temporal seqüencial foi apresentada com um maior decréscimo entre os períodos T2 e T3. Desta forma, caracterizando uma queda mais brusca na média do grupo.

4.1.3 Número de Artigos Publicados em Periódicos

Esta seção faz uma análise dos dados de cada pesquisador levando em consideração o número de artigos publicados em periódicos. Pretende-se identificar se os pesquisadores evoluíram com relação a sua produção. As Figuras 4.11, 4.12, 4.13, 4.14 e 4.15 apresentam o resultado da análise, no qual se identifica o número de artigos publicados em eventos (em uma escala de 0 a 25) dos 228 pesquisadores (P1 ao P228) em 3 períodos de tempo (T1, T2 e T3).

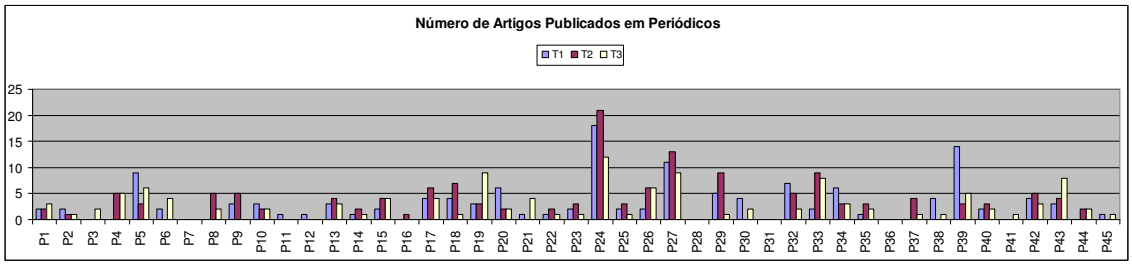

Figura 4.11: Número de artigos publicados em periódicos de P1 à P45.

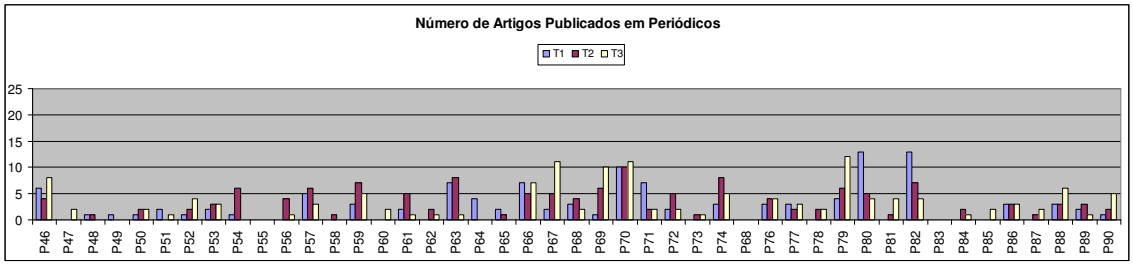

Figura 4.12: Número de artigos publicados em periódicos de P46 à P90.

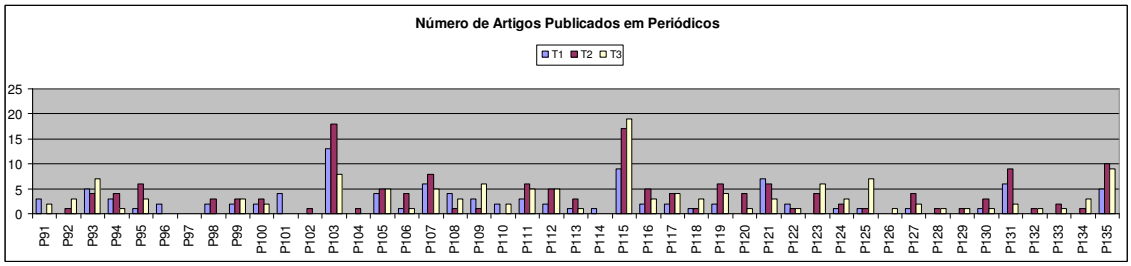

Figura 4.13: Número de artigos publicados em periódicos de P91 à P135.

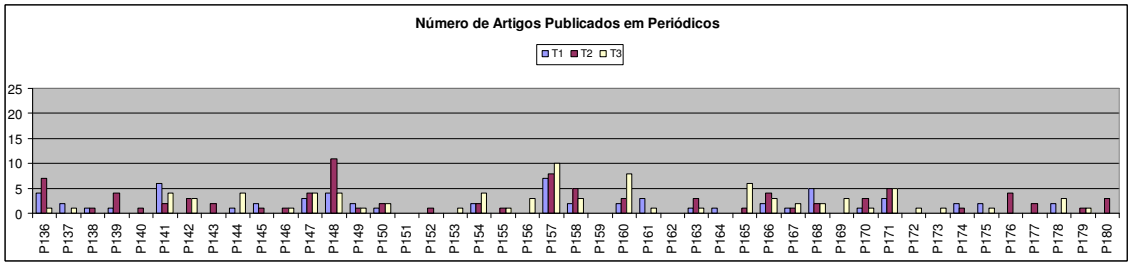

Figura 4.14: Número de artigos publicados em periódicos de P136 à P180.

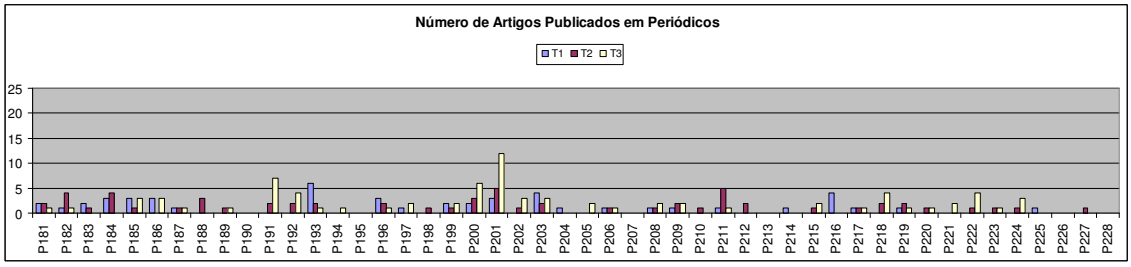

Figura 4.15: Número de artigos publicados em periódicos de P181 à P228.

Os pesquisadores do G1, em sua maioria, confirmaram sua senioridade na pesquisa apresentando uma produção mais alta que a do G2 e principalmente do G3, analisando-se os três períodos de tempo. Mas, se analisarmos o número de trabalhos de alguns pesquisadores do G3, pode-se perceber que estes aumentaram sua produção, no tempo T3, com relação aos pesquisadores do G1 do G2.

Com relação à evolução na passagem seqüencial pelos 3 períodos de tempo, 51 pesquisadores do G1, 40 pesquisadores do G2 e 10 pesquisadores do G3 apresentaram aumento de sua produção. Com relação ao decréscimo da produção, nos 3 períodos de tempo, 43 pesquisadores do G1, 19 do G2 e 5do G3 diminuíram suas publicações em periódicos. Se analisarmos períodos não seqüenciais, alguns pesquisadores do G1 e do G2 também apresentaram decréscimo em determinados períodos de tempo. Já os pesquisadores do G3, comparando-se os 3 períodos de tempo, nenhum apresentou decréscimo na produção.

A média de artigos publicados em periódicos dos pesquisadores do G2 e do G3 apresentou evolução em todos os períodos analisados. Já a média dos pesquisadores do G1 apresentou redução de produção na passagem de T2 para T3, conforme os dados a seguir:

- média do G1 foi 2.82 em T1, 3.38 em T2 e 2.99 em T3;
- média do G2 foi 1.38 em T1, 2.04 em T2 e 2.15 em T3;
- média do G3 foi 0.65 em T1, 0.96 em T2 e 1.15 em T3.

Na análise realizada para artigos publicados em periódicos identificou-se que, na média, os pesquisadores apresentaram evolução positiva em todos os períodos de tempo analisados. Somente o G1, no tempo T2 para T3 apresentou redução de produção. Este resultado pode ter ocorrido em função do alto número de pesquisadores que tiveram um elevado número de artigos publicados em T2, mas em T3 diminuíram abruptamente sua produção. Também pode-se vincular este fato à queda já apresentada por este grupo neste mesmo período de tempo, na produção de artigos publicados em conferências. Esta queda de produção é muito perceptível em doutorandos de instituições estrangeiras de renome que publicavam no contexto de grupos de pesquisa consolidados. Em seu retorno a suas universidades de origem, a produção passa a ser mais individualizada.

O aumento de produção ocorrido de um modo geral pode ter relação com o aumento de seu número de co-autores, conforme apresentado na Figura 1. Esta análise permite evidenciar a importância da constituição de projetos de pesquisa de maior porte e multi-institucionais. Destaca-se que alguns pesquisadores dos 3 grupos publicaram um número de artigos muito superior aos demais pares, o que influenciou no resultado de aumento de produção se analisada a média entre todos os grupos.

4.1.4 Número de Dissertações Orientadas

Esta seção faz uma análise dos dados de cada pesquisador levando em consideração o número de dissertações orientadas concluídas e em andamento. Pretende-se identificar se os pesquisadores evoluíram com relação a sua rede de orientações. As Figuras 4.16, 4.17, 4.18, 4.19, 4.20 apresentam o resultado das análises, no qual se identifica o número de dissertações orientadas concluídas e em andamento (em uma escala de 0 a 30) dos 228 pesquisadores (P1 ao P228) em 3 períodos de tempo (T1, T2 e T3).

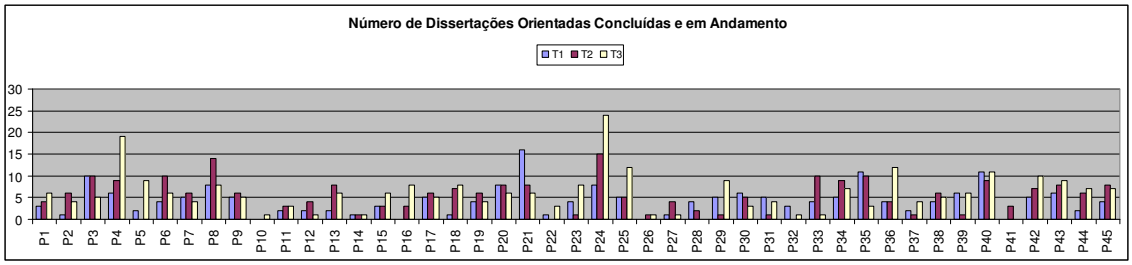

Figura 4.16: Número de dissertações orientadas de P1 à P45.

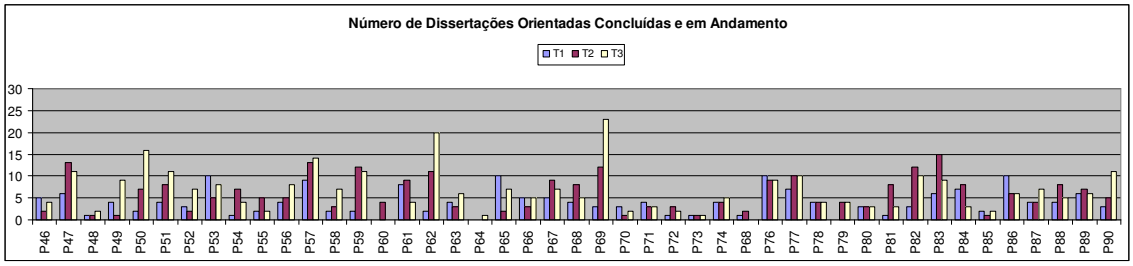

Figura 4.17: Número de dissertações orientadas de P46 à P90.

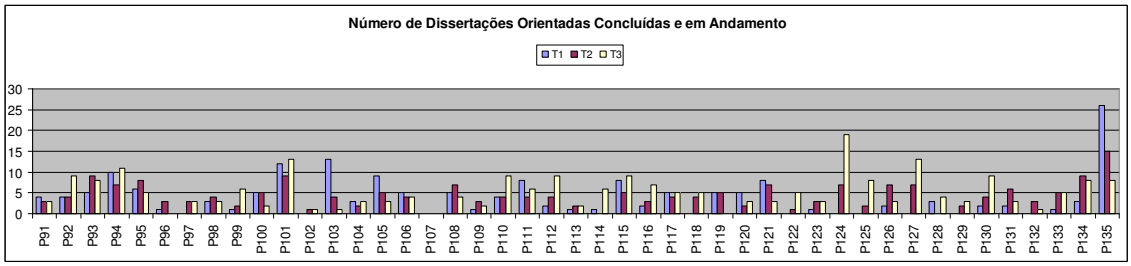

Figura 4.18: Número de dissertações orientadas de P91 à P135.

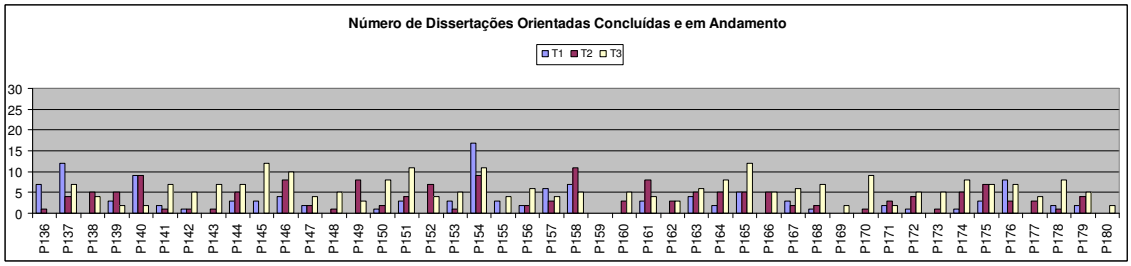

Figura 4.19: Número de dissertações orientadas de P136 à P180.

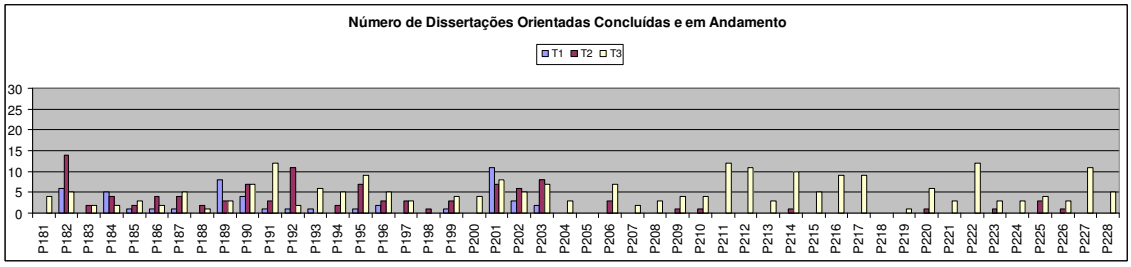

Figura 4.20: Número de dissertações orientadas de P181 à P228.

Os pesquisadores do G1, em sua maioria, têm um número de orientações superior aos demais grupos analisados, principalmente nos períodos T1 e T2. Isso deve-se a sua senioridade e tempo de pesquisa maior que os demais grupos. Com relação à evolução positiva na passagem seqüencial pelos 3 períodos de tempo, 65 pesquisadores do G1, 64 pesquisadores do G2 e 24 pesquisadores do G3 apresentaram aumento no número de orientações. Com relação à evolução negativa nos 3 períodos de tempo, 33 pesquisadores do G1 e 12 do G2 diminuíram suas orientações. Se analisarmos períodos não seqüenciais, apenas 8 pesquisadores do G1 apresentaram redução de orientações em determinados períodos de tempo e nenhum pesquisador do G2 e G3 apresentaram redução de orientações nos 3 períodos de tempo.

A média de orientações de mestrado dos pesquisadores dos 3 grupos apresentou evolução positiva, conforme os dados a seguir:

- média do G1 foi 4.21 em T1, 5.20 em T2 e 5.88 em T3;
- média do G2 foi 2.62 em T1, 3.94 em T2 e 5.46 em T3;
- média do G3 foi 0.08 em T1, 0.77 em T2 e 5.38 em T3.

Na análise realizada para orientações de mestrado identificou-se que, na média, os pesquisadores apresentaram evolução positiva em todos os períodos analisados. Destaca-se que os pesquisadores do G2 e principalmente do G3, na passagem de T2 para T3 evoluíram muito em números absolutos. Isto se deve ao fato de alguns pesquisadores destes grupos terem orientado um alto número de alunos, o que culminou em um aumento significativo da média. Ao mesmo tempo alguns pesquisadores de todos os grupos não possuem nenhuma orientação. Como a maioria dos pesquisadores do G3 concluíram o doutorado há pouco tempo, seu número de orientações nos períodos T1 e T2 é quase nulo, o que já era esperado. Já no período T3 eles obtiveram um número de orientações elevado, se aproximando dos pesquisadores do G1 e G2. Isto demonstra que estão participando ativamente do processo de orientação.

Outra análise interessante é com relação aos pesquisadores do G1. Eles mantiveram o número de orientações de dissertação dentro de uma mesma faixa nos 3 períodos analisados, o que configurou um aumento na média geral do grupo não tão acentuado como o ocorrido para os outros grupos.

4.1.5 Número de Teses Orientadas

Esta seção faz uma análise dos dados de cada pesquisador levando em consideração o número de teses orientadas concluídas e em andamento. Pretende-se identificar se os pesquisadores evoluíram com relação a sua rede de orientações. As Figuras 4.21, 4.22, 4.23, 4.24 e 4.25 apresentam o resultado da análise, no qual se identifica o número de teses orientadas concluídas e em andamento (em uma escala de 0 a 15) dos 228 pesquisadores (P1 ao P228) em 3 períodos de tempo (T1, T2 e T3).

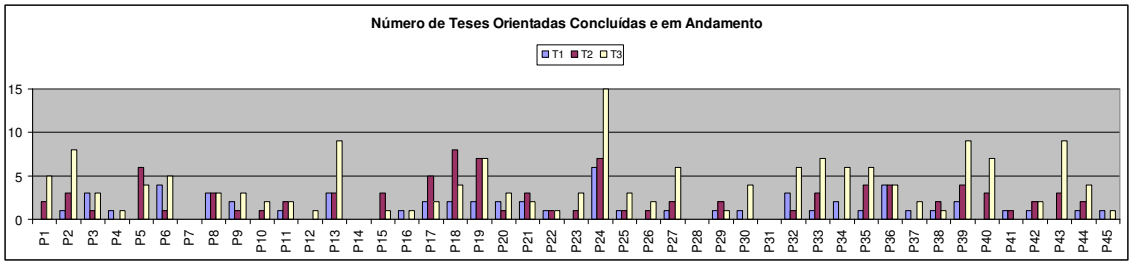

Figura 4.21: Número de teses orientadas de P1 à P45.

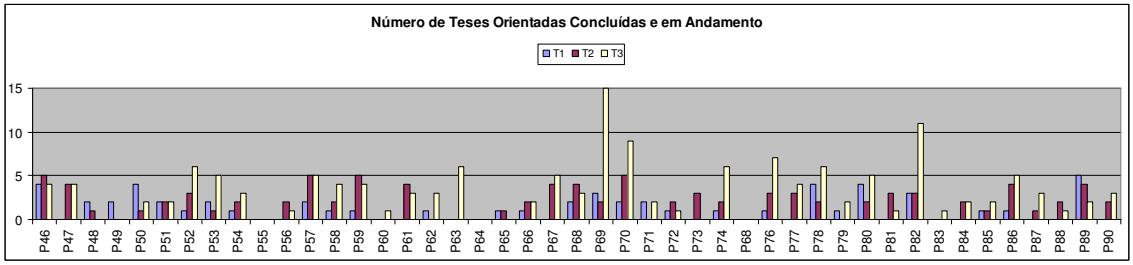

Figura 4.22: Número de teses orientadas de P46 à P90.

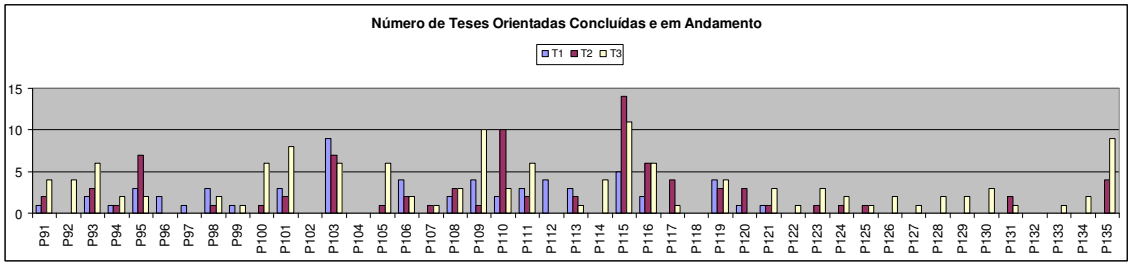

Figura 4.23: Número de teses orientadas de P91 à P135.

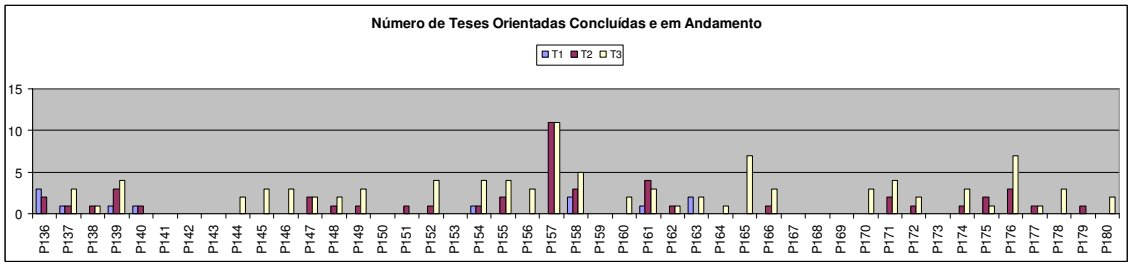

Figura 4.24: Número de teses orientadas de P136 à P180.

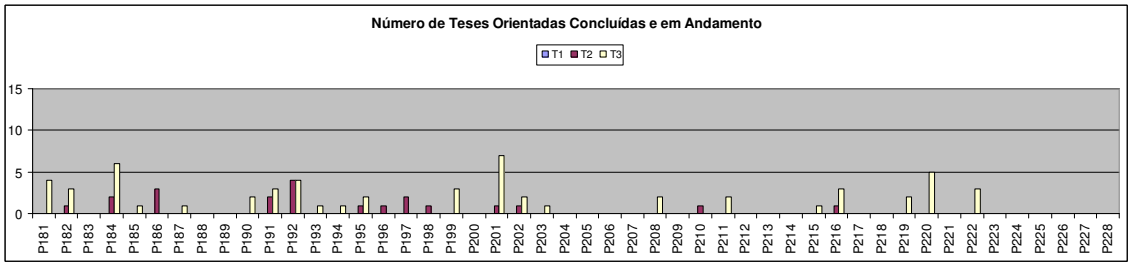

Figura 4.25: Número de teses orientadas de P181 à P228.

Os pesquisadores do G1 apresentam um número de orientações superior aos demais grupos analisados, nos 3 períodos de tempo. Com relação à evolução temporal, 75 pesquisadores do G1, 57 do G2 e 8 do G3 apresentaram aumento no número de orientações. Com relação à redução da produção nos 3 períodos, 14 pesquisadores do G1 e 2 do G2 apresentaram uma diminuição nas orientações de doutorado. Se analisarmos períodos não seqüenciais, alguns 10 pesquisadores do G1 e 1 do G2 apresentaram oscilações na produção em determinados períodos. Nenhum pesquisador do G3 apresentou diminuição de orientações na análise seqüencial, nem redução em determinados períodos de tempo.

Nenhum pesquisador do G3 possui orientações de doutorado no período T1. Apenas 8 pesquisadores do G2 apresentaram orientação neste mesmo período. No período T2, apenas 2 pesquisadores do G3 possuem orientações de doutorado. Isto se justifica pelo fato de terem defendido o doutorado muito próximo ou até durante os períodos de avaliação realizados em T1 e T2. Dos dois pesquisadores do G3 que possuem orientação em T2, apenas 1 apresentou evolução positiva no período T3. Já 10 pesquisadores do G1, 15 pesquisadores do G2 e 17 pesquisadores do G3 não orientaram nenhuma tese de doutorado em nenhum dos 3 períodos de tempo analisados. Levando-se em conta a diferença do tamanho de cada um dos grupos analisados, pode-se perceber que a senioridade teve forte influência no percentual de pesquisadores sem orientação encontrados.

A média de orientações de doutorado dos pesquisadores dos 3 grupos apresentou a evolução apresentada a seguir:

- média do G1 foi 1.45 em T1, 2.19 em T2 e 3.45 em T3;
- média do G2 foi 0.15 em T1, 0.94 em T2 e 2.09 em T3;
- média do G3 foi 0 em T1, 0.08 em T2 e 0.73 em T3.

Na análise realizada para orientações de doutorado identificou-se que, na média, os pesquisadores apresentaram evolução em todos os períodos de tempo analisados. Destaca-se que muitos pesquisadores do G2 e do G3 não orientaram nenhuma tese de doutorado, apesar de pertencerem a programa de pós-graduação Stricto Sensu. Sendo isso mais uma vez explicado pelo pouco tempo de conclusão do doutorado dos pesquisadores destes grupos.

4.2 Agrupamento

O objetivo desta seção é apresentar experimentos envolvendo uma das técnicas de mineração de dados, algoritmos de agrupamento. Pretende-se mostrar utilizando os experimentos de agrupamentos usando como elemento de comparação as variáveis co-autores, produção e orientação se os grupos definidos manualmente na seção 3.2 correspondem aos agrupamentos gerados pelo algoritmo `Simple K-Means`[11].

4.2.1 Agrupamento por Co-autores

A Figura 4.26 ilustra o resultado do experimento onde utilizamos o número de co-autores dos 228 pesquisadores como elemento de comparação, considerando os 3 grupos definidos na seção 3.2.

[11] Algoritmo não supervisionado que fornece uma classificação automática baseada em análise e comparações entre os valores numéricos dos dados.

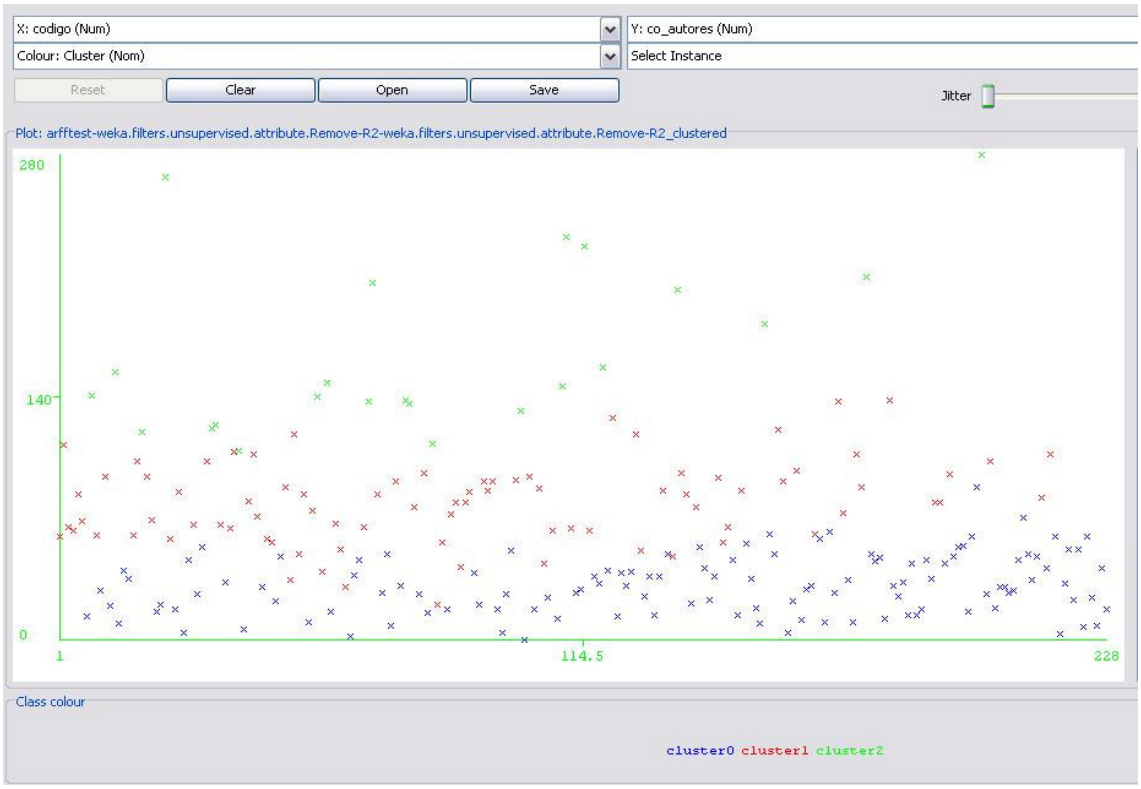

Figura 4.26: Agrupamento por co-autores.

Observa-se na Figura 4.26 que os 3 grupos de pesquisadores definidos na seção 3.1 não apresentam semelhança com os 3 agrupamentos encontrados no experimento. Isto mostra que o tempo de conclusão de doutorado não tem influência no número de co-autores destes pesquisadores. Isto nega a hipótese de que a experiência do pesquisador não é determinante no número de co-autores.

4.2.2 Agrupamento por produção

A Figura 4.27 ilustra o resultado do experimento onde utilizamos como elemento de comparação a produção dos 228 pesquisadores, ou seja, o número de artigos publicados em periódicos e o número de trabalhos publicados em eventos, considerando os 3 grupos definidos na seção 3.2.

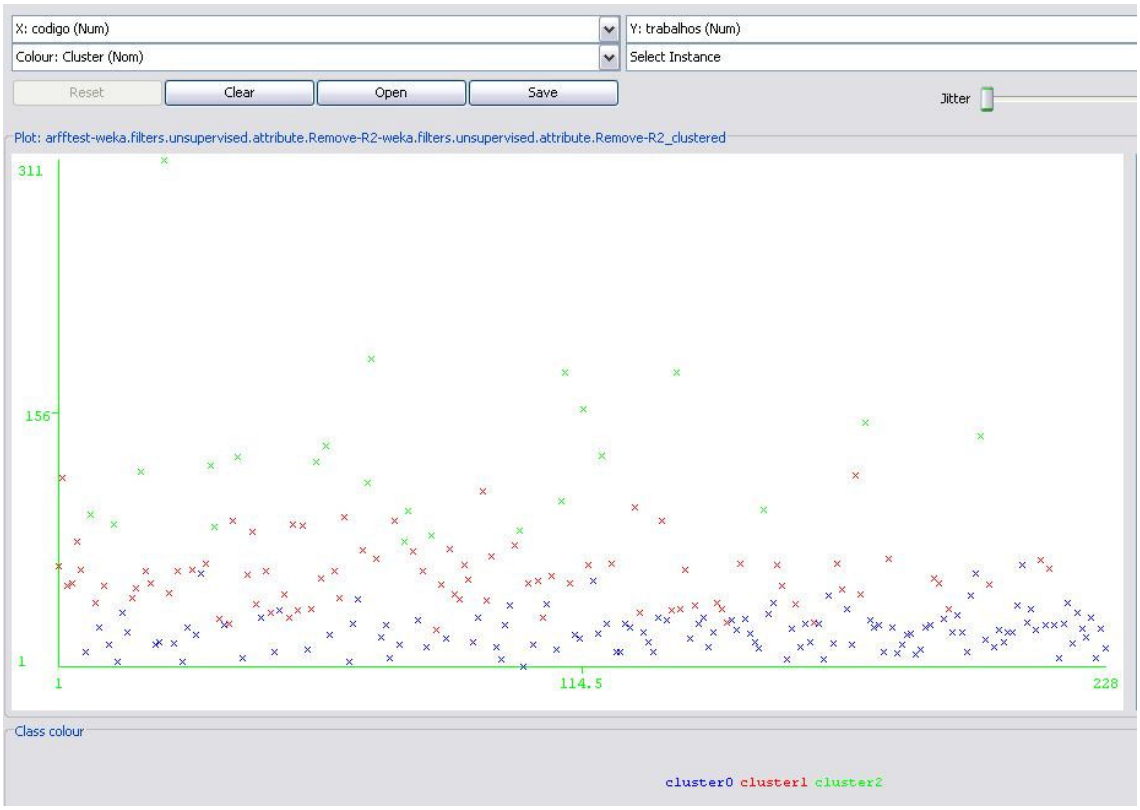

Figura 4.27: Agrupamento por produção.

Observa-se na Figura 4.27 que os 3 grupos de pesquisadores definidos na seção 3.1 não apresentam semelhança com os 3 agrupamentos encontrados no experimento. Isso também mostra que o tempo de conclusão do doutorado não tem influência no número de publicações dos pesquisadores, uma vez que pesquisadores do G3 publicaram mais que pesquisadores do G1 e G2. Destaca-se que alguns dos pesquisadores dos grupos G2 e G3 possuem produção muito próxima de alguns pesquisadores do G1.

4.2.3 Agrupamento por orientações

A Figura 4.28 ilustra o resultado do experimento onde utilizamos como elemento de comparação as orientações dos 228 pesquisadores, ou seja, o número de orientações de mestrado e de doutorado, considerando os 3 grupos definidos na seção 3.2.

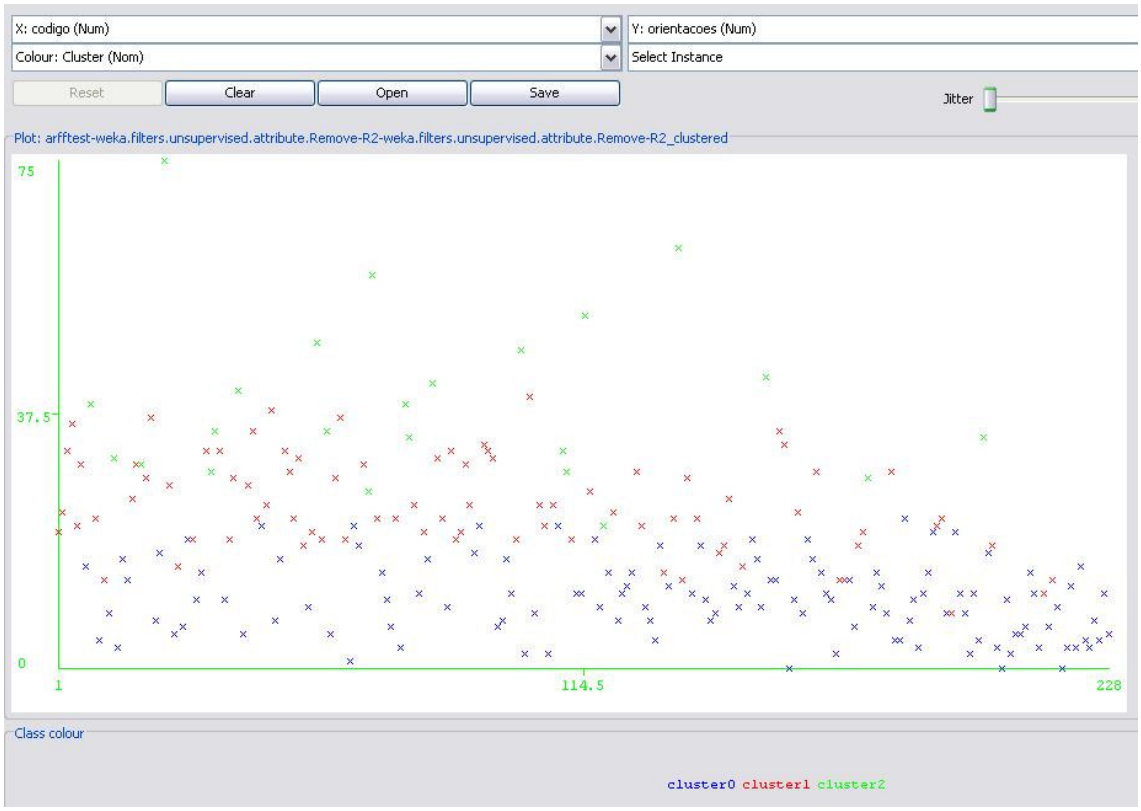

Figura 4.28: Agrupamento por orientações.

Observa-se na Figura 4.28 que os 3 grupos de pesquisadores definidos na seção 3.2 não apresentam semelhança com os 3 agrupamentos encontrados no experimento. Isto mostra que os pesquisadores do grupo G1 possuem um número maior de orientações do que a grande maioria dos pesquisadores dos grupos G2 e G3. Evidencia-se, ainda, que alguns pesquisadores dos grupos G3 e G2 aumentaram significativamente suas orientações, agrupando-se com muitos dos pesquisadores experientes do G1.

4.3 Regressão

O objetivo desta seção é apresentar experimentos envolvendo outra técnica de mineração de dados, algoritmos de regressão. Para os experimentos com regressão, utilizamos regressão linear simples para descobrir o número da produção dos pesquisadores a partir do número de orientações, analisando sua correlação em comparação com a realidade. Os experimentos com regressão linear múltipla tiveram como objetivo descobrir o número da produção dos pesquisadores a partir do número de co-autores juntamente com o número de orientações de mestrado e doutorado, também analisando sua correlação em comparação com a realidade. O algoritmo utilizado para regressão linear foi o `Regressn`[12].

4.3.1 Regressão Linear Simples

O objetivo deste experimento é descobrir o número da produção dos pesquisadores a partir do número de orientações, analisando sua correlação em comparação com a

[12] É utilizado para relação de previsão, onde calcula uma relação linear entre uma variável dependente e uma independente.

realidade. As Figuras 4.29, 4.30, 4.31, 4.32, 4.33 apresentam o resultado do experimento realizado com os 228 pesquisadores. Foi utilizado o número de artigos publicados em periódicos somado com o número de trabalhos publicados em eventos como variável dependente. O número de orientações de mestrado e de doutorado foi utilizado como variável independente.

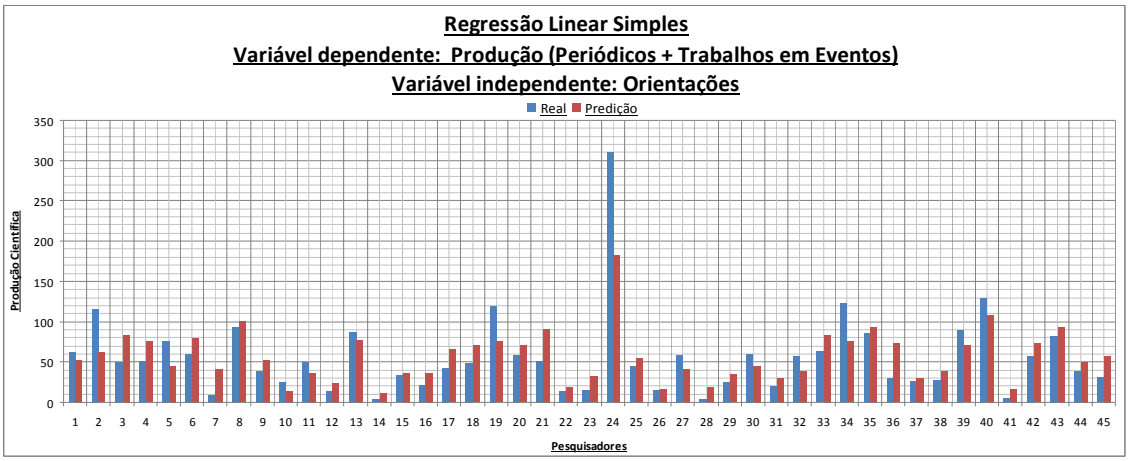

Figura 4.29: Regressão Linear Simples de P1 à P45.

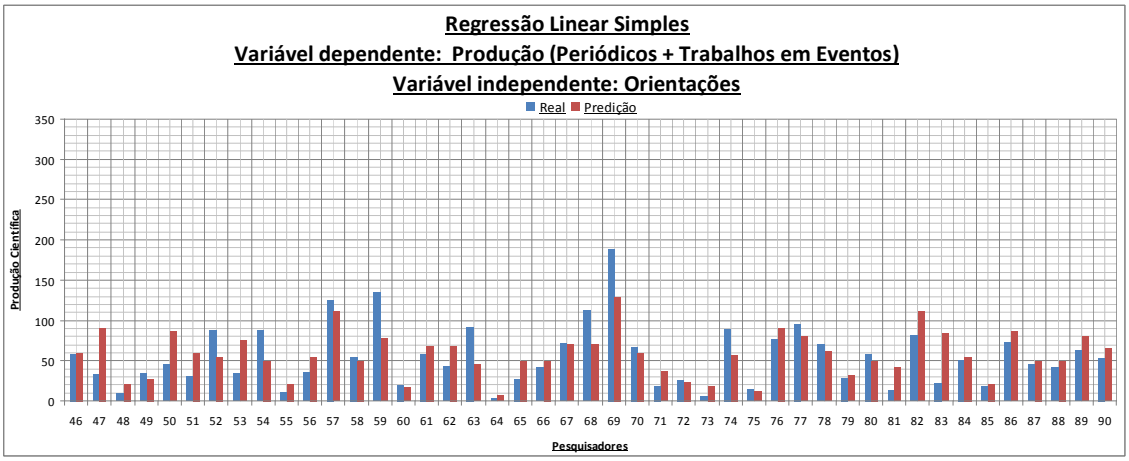

Figura 4.30: Regressão Linear Simples de P46 à P90.

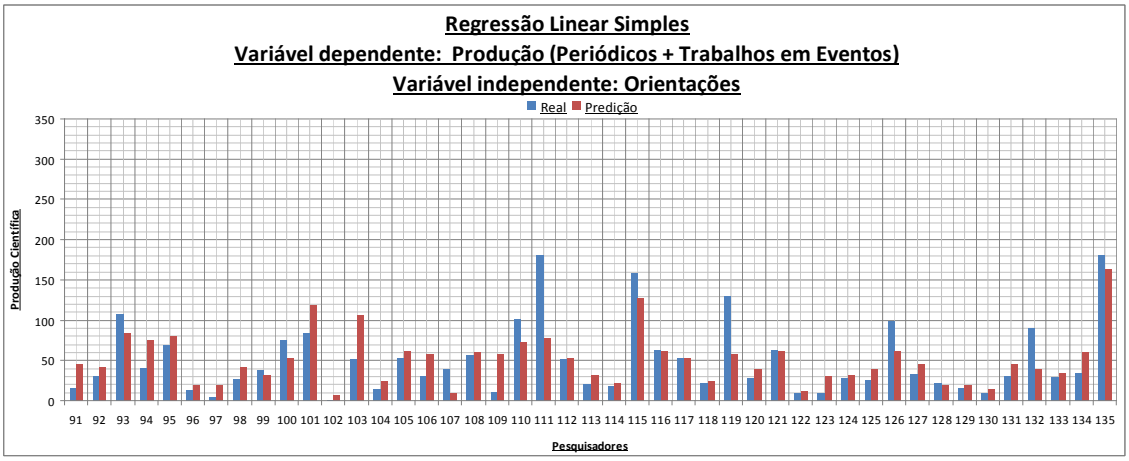

Figura 4.31: Regressão Linear Simples de P91 à P135.

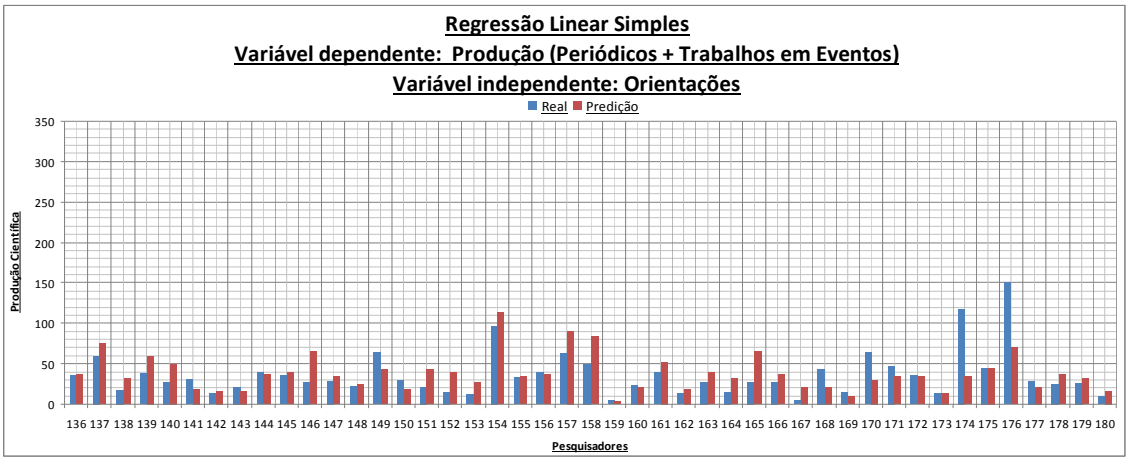

Figura 4.32: Regressão Linear Simples de P136 à P180.

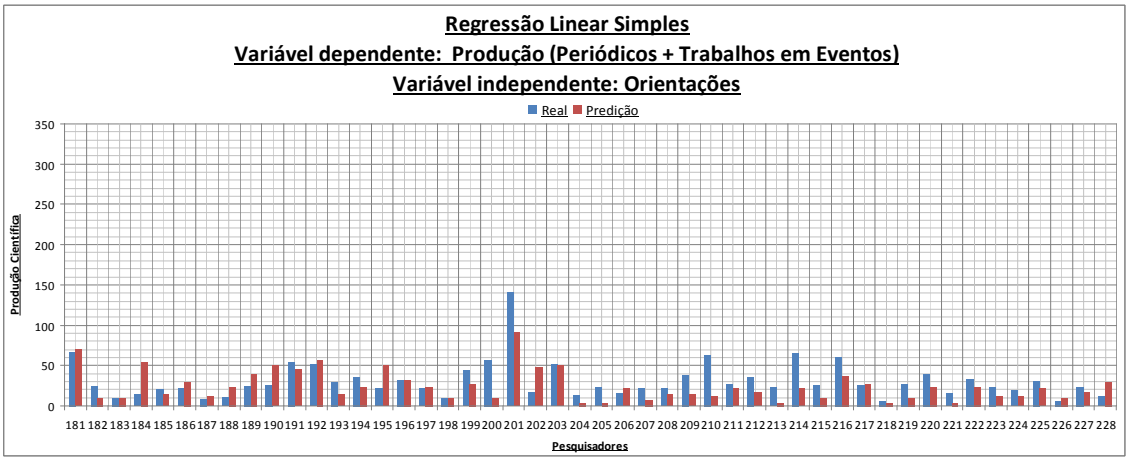

Figura 4.33: Regressão Linear Simples de P181 à P228.

Observa-se nas Figuras 4.29, 4.30, 4.31, 4.32 e 4.33 que, para a grande maioria dos pesquisadores, independente do grupo em que estão, o número da produção real (barra da esquerda) corresponde ao número da predição de sua produção (barra da direita). Mesmo assim, ainda tivemos casos de pesquisadores que diferença algumas vezes chegou a dobrar, seja para mais ou para menos. Para esse experimento de regressão linear simples foi encontrado um coeficiente de correlação[13] igual a 0.75639, o que indica uma correlação forte[14] para a variável orientação. Pode-se identificar, através destas diferenças, que o número de orientações tem uma considerável e relevante influência no impacto das publicações do pesquisador. O fato de mesmo tendo uma correlação forte, estar próximo ao limítrofe da correlação moderada explica os resultados diversos apresentados, com alguns se aproximando bastante da realidade e quanto outros fugindo bastante disso.

[13] Coeficiente de correlação: Indica a relação entre a produção do pesquisador e a soma dos diferentes pesos das variáveis independentes, podendo variar de -1 a +1, o que representa uma escala que vai de menos significativo até mais significativo.

[14] Interpretação do coeficiente de correlação: de 0.00 a 0.19 indica uma correlação bem fraca, de 0.20 a 0.39 correlação fraca, 0.40 a 0.69 correlação moderada, de 0.70 a 0.89 correlação forte e de 0.90 a 1.00 correlação muito forte.

4.3.2 Regressão Linear Múltipla

O objetivo deste experimento é descobrir o número da produção dos pesquisadores a partir do número de co-autores e o número de orientações de mestrado e de doutorado, analisando sua correlação em comparação com a realidade. As Figuras 4.34, 4.35, 4.36, 4.37 e 4.38 ilustram o resultado do experimento realizado com os 228 pesquisadores. Foi utilizado o número de artigos publicados em periódicos somado com o número de trabalhos publicados em eventos dos 228 pesquisadores como variável dependente. O número de co-autores juntamente com o número de orientações de mestrado e de doutorado foram utilizados como variáveis independentes.

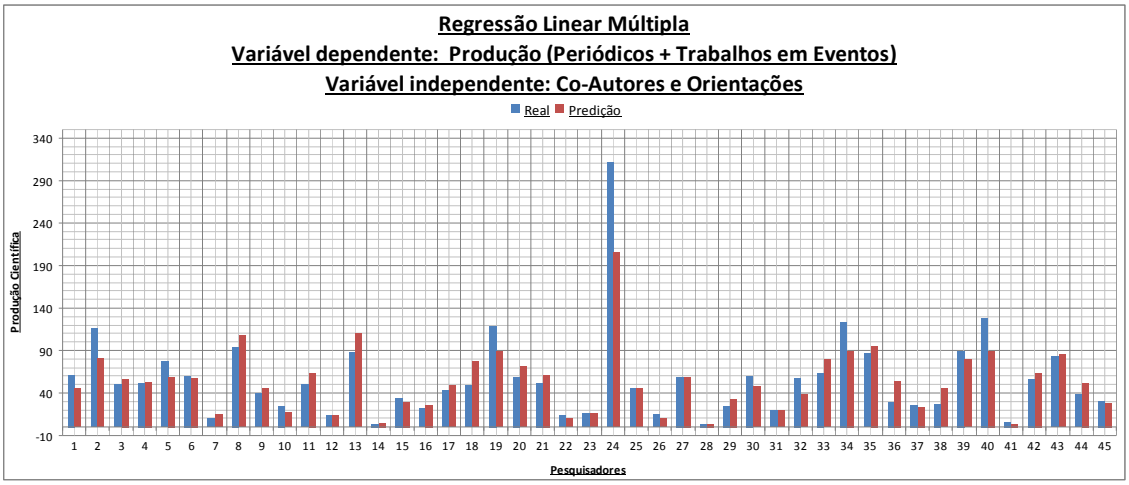

Figura 4.34: Regressão Linear Múltipla de P1 à P45.

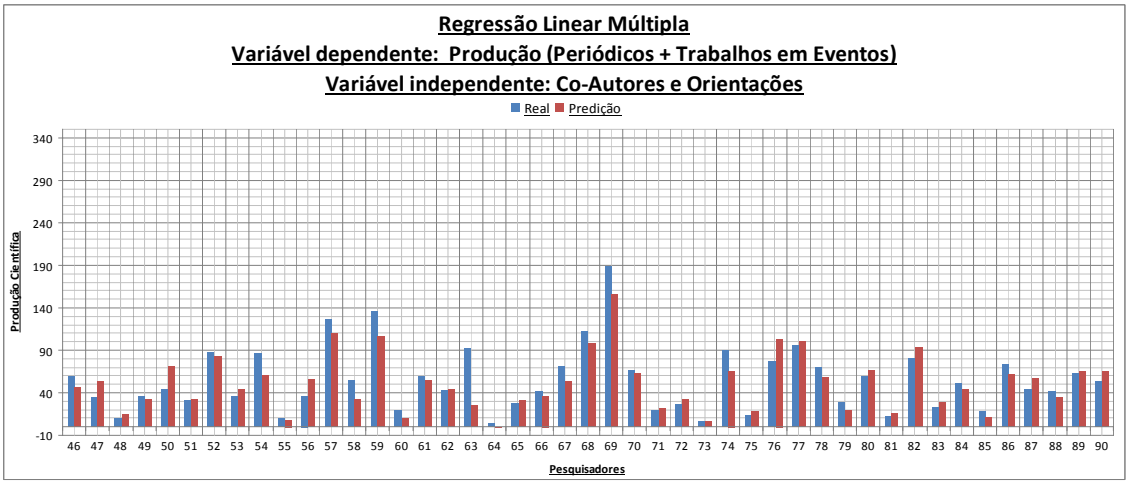

Figura 4.35: Regressão Linear Múltipla de P46 à P90.

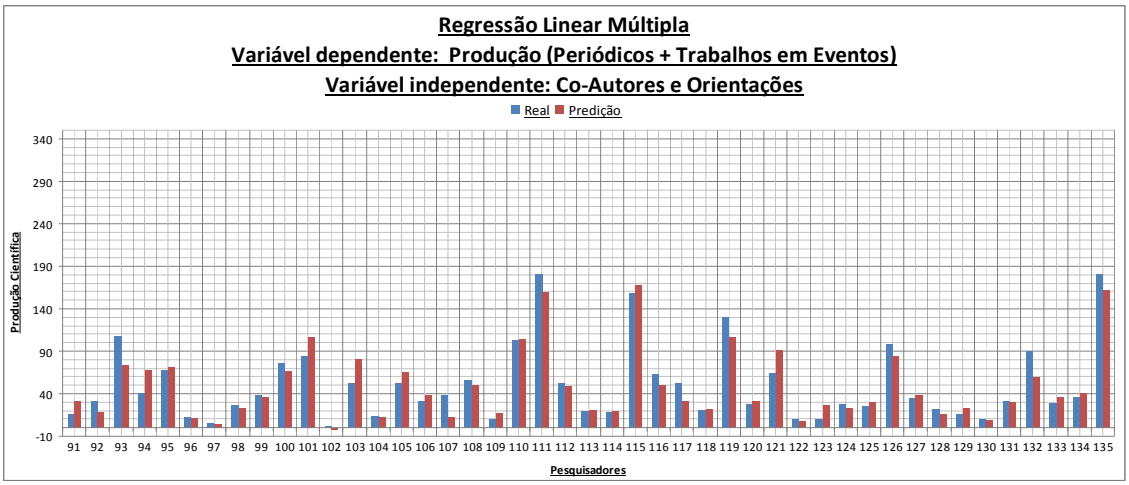

Figura 4.36: Regressão Linear Múltipla de P91 à P135.

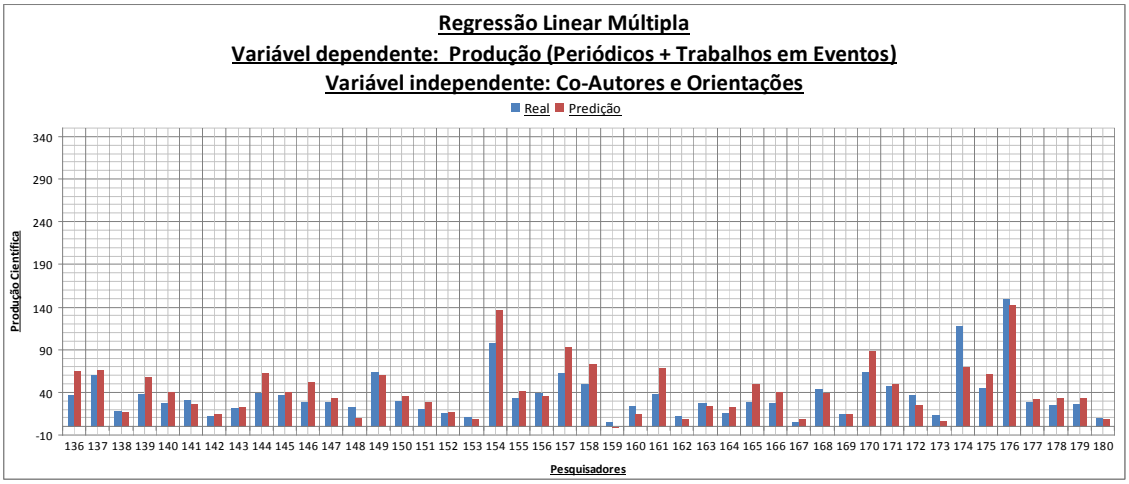

Figura 4.37: Regressão Linear Múltipla de P136 à P180.

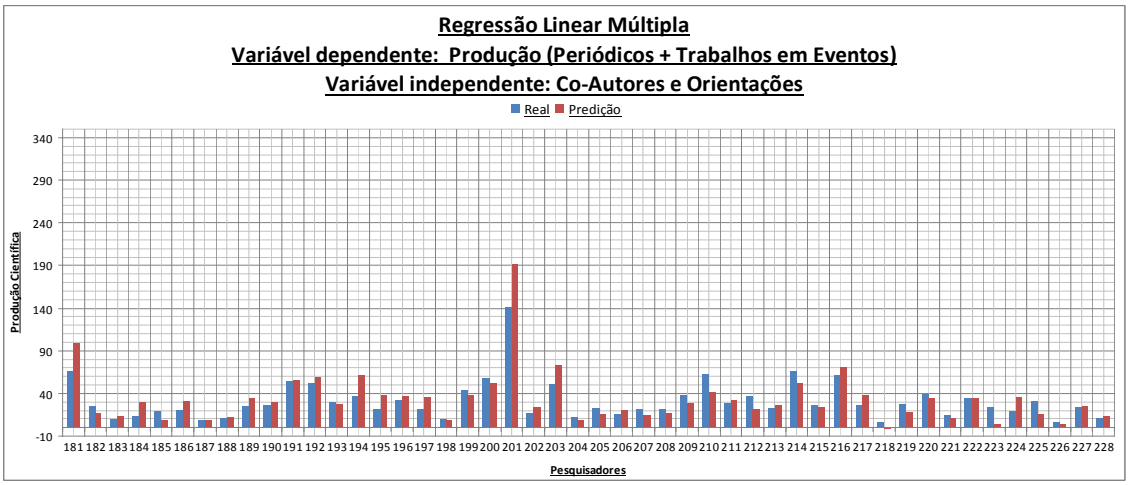

Figura 4.38: Regressão Linear Múltipla de P181 à P228.

Observa-se nas Figuras 4.34, 4.35, 4.36, 4.37 e 4.38 que o número da produção real dos pesquisadores (barra da esquerda) é compatível com o número da predição de sua produção (barra da direita). Em comparação com a regressão linear simples, em poucos casos há uma discrepância entre os valores, como exemplificado em P24, P40, P58,

P85, P146, P174 e P184, onde houve uma diferença muito significativa de incompatibilidade entre o real e a predição. Isto pode ser explicado pelo alto número de produção destes pesquisadores, o que pode ter influenciado no conjunto de dados utilizados. Contudo, houve casos em que a predição foi negativa, como em P64, P102, P159 e P218. Dessa forma, pode-se concluir que o número de co-autores de um pesquisador, juntamente com o número de orientações tem impacto relevante nos números de sua produção, pois estes pesquisadores tiveram um fraco desempenho levando-se em conta aspectos de co-autoria, orientações e produção. Isso fica claro na análise do gráfico onde as barras identificando a realidade e a predição ficam muito próximas umas da outras para a quase totalidade dos pesquisadores analisados.

No experimento de regressão linear múltipla foi encontrado um coeficiente de correlação igual a 0.90560, o que indica uma correlação muito forte. Isto evidencia que houve uma melhora muito significativa quando as variáveis co-autores e orientação foram incorporadas ao modelo. Nesse caso, a predição se aproximou muito do valor real. Percebe-se esta melhora quando se compara a maioria dos pesquisadores nos resultados dos experimentos utilizando-se regressão linear simples e, após, com regressão linear múltipla.

5 CONCLUSÃO

O trabalho apresentou importantes análises de produção científica através da aplicação de experimentos utilizando métodos de descoberta de conhecimento em base de dados (KDD) e mineração dados, como: evolução temporal, algoritmos de agrupamento e algoritmos de regressão. Tais experimentos foram aplicados em uma base de dados composta pelas informações curriculares dos pesquisadores extraídas da Plataforma Lattes.

Os experimentos apresentaram uma contribuição muito grande em termos de observação de aspectos relevantes a um pesquisador para seu resultado de produção. Foi possível verificar após a analise temporal que o tempo de pesquisa é um fator bastante determinante para o número de co-autores, produção de artigos seja em periódicos ou em eventos, orientações de dissertações e teses, apresentados por um pesquisador, comprovando a senioridade ser bastante relevante para tal fato. A posição de crescimento quase que exponencial apresentadas nesses pontos por pesquisadores mais jovens, volta a incrementar a importância do tempo de pesquisa para os resultados encontrados, levando-nos a crer que a evolução dos resultados é contínua e diretamente proporcional ao aumento do tempo de pesquisa. Contudo, após os testes com algoritmos de agrupamento, verificou-se que o tempo de conclusão de doutorado não é fator determinante para o resultado encontrado de número de co-autores, produção e orientações. Isso pode nos mostrar que os algoritmos de agrupamentos da forma como foram configurados e executados não apresentam resultados condizentes com a realidade. Nos resultados apresentados pelos testes com regressão, observou-se que quanto mais variáveis forem utilizadas, mais o resultado pode ser previsto e igualado ao que encontramos na realidade.

Os experimentos ainda mostram que para termos uma opinião e um resultado mais embasado, não basta apenas uma análise simples. É preciso a combinação de análises de diversos tipos envolvendo mais de uma técnica conhecida para que obtenhamos uma resolução mais próxima da verdade.

Como trabalhos futuros, pretende-se identificar como critério de agrupamento não o tempo de conclusão de doutorado, mas sim o tempo em que o pesquisador ingressou em um programa de pós-graduação *Stricto Sensu*. Além disto, outra possibilidade é de incorporar mais elementos para a realização dos experimentos. Isto pode ser possível definindo-se um modelo de perfil para os pesquisadores, formado por diversos atributos que possam representar desde a área de atuação de um pesquisador, as citações para seus artigos, se ele é membro de comitê de programa de conferência ou revisor de periódico, bem como se possui livros ou capítulos de livros publicados. Estes elementos podem ser importantes ao se analisar a carreira científica de um pesquisador, uma vez que as informações geradas podem ser relevantes para auxiliar Agências, Universidades e Entidades no processo de tomada de decisão.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. VLDB, p. 487-499. 1994.

ARRUDA, Denis., BEZERRA, F., NERIS, V. A., ROCHA de TORO, P., WAINER, J. Brazilian Computer Science Research: Gender and Regional Distributions. Scientometrics, 2009.

BRADLEY, Neil. The XML companion. 3.ed. Boston: Pearson Education, 2002.

BRUSSO, Marcos José. Access Miner: uma proposta para a extração de regras de associação aplicada à mineração do uso da web. 2000. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

CAPES. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Disponível em <http://www.capes.gov.br>. Acesso em 20 set. 2009.

CAZELLA, Sílvio C. Aplicando a relevância da opinião de usuários em sistema de recomendação para pesquisadores. 2006. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

CNPQ. Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em: <http://www.cnpq.br/cnpq/index.htm>. Acesso em: 10 nov. 2009.

DÉCIO, Otávio C. XML: guia de consulta rápida. São Paulo: Novatec, 2000.

EIBE, Frank et al. Weka: A Machine Learning Workbench for Data Mining. In: _____. The Data Mining and Knowledge Discovery Handbook. New York: Springer, 2005. P. 1305-1314.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI Magazine. 1996. p. 37-54.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. Knowledge Discovery in Databases: An overview. In: AI Magazine. American Association for Artificial Intelligence. Menlo Park, 1992. p. 57-70.

FREITAS, Juliana Gonçalves de. Uma ferramenta para clusterização de perfis de usuários baseada em dados qualitativos. 2005. Trabalho de Conclusão de Curso (Graduação em Análise de Sistemas) - Universidade do Vale do Rio dos Sinos, São Leopoldo, 2005.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel: Data mining: um guia prático. Rio de Janeiro: Editora Campus, 2005.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: a robust clustering algorithm for categorical attributes. In: 15th Int'l Conf. on Data Engineering (ICDE). Sydney, Australia. 1999. p. 512-521.

HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. Burnaby, British Columbia, Canada: Morgam Kaufmam Publishes, Inc., 2001.

HANNEL, Kelly. Qualificação de pesquisadores por área da ciência da computação com base em uma ontologia de perfil. 2008. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

HUANG, Zhexue. Extensions to the k-means algorithm for clustering large data sets with categorical values. In: Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Hingham, MA, USA, v. 2, p. 283-304. 1998.

IDAMS. Internationally Developed Data Analysis and Management Software
Disponível em: <http://portal.unesco.org/ci/en/ev.php-URL_ID=2070&URL_DO=DO_TOPIC&URL_SECTION=201.html>. Acesso em 15 nov. 2009.

JOHN, George H. Enhancements To The Data Mining Process. Phd. Dissertation – Stanford University. Phd. Dissertation, 1997.

LATTES. Plataforma Lattes. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 5 nov. 2009.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, p.281-297. 1967.

MATHWORKS "Statistics Toolbox: K-means clustering". Disponível em: <http://www.mathworks.com/access/helpdesk/help/toolbox/stats/kmeans.html>. Acesso em: 01 dez. 2009.

NATARAJAN, Rajesh; SHEKAR, B. Interestingness of association rules in data mining: issues relevant to e-commerce. India: Sadhana, 2005. p. 291-309.

OLIVEIRA, Eder de; SOUZA BERMEJO, Paulo Henrique de; MEDINA KERN, Vinícius. GeraLattes: extração de informação gerencial de currículos de pesquisadores usando XML. In: WORKSHOP DE COMPUTAÇÃO DA REGIÃO SUL, 1, 2004, Florianópolis, Anais. Florianópolis: Unisul, 2004.

QUINLAN, J. Ross. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

RIBEIRO JUNIOR, Luiz Carlos et al. Identificação de Áreas de Interesse a partir da Extração de Informações de Currículos Lattes/XML. In: Escola Regional de Banco de Dados, 2005, Porto Alegre. Escola Regional de Banco de Dados - ERBD, 2005.

ROMÃO, Wesley. Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia. 2002. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2002.

SILVA, M. P. S. Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka. Livro da Escola Regional de Informática Rio de Janeiro - Espírito Santo. Porto Alegre: Sociedade Brasileira de Computação, 2004, v. 1, p. 1-20.

TOMAZELA, Bruno; CIFERRI, C. D. A.; TRAINA JR, Caetano. Reconciliando Dados de Cunho Acadêmico. In: XXIII Simpósio Brasileiro de Banco de Dados, 2008, Campinas. Anais do XXIII Simpósio Brasileiro de Banco de Dados, 2008. p. 283-297.

TWO CROWS CORPORATION. Introduction to data mining and knowledge discovery. 3 ed. 2005. Disponível em: <http://www.twocrows.com/intro-dm.pdf>. Acesso em: 15 set. 2009.

VELOSO, Renê Rodrigues. Java e XML: processamento de documentos XML com Java. 2. ed. São Paulo: Novatec, 2007.

VOLTOLINI, Richardson F. Discretização e geração de dados em aprendizado de máquina. 2006. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2006.

WAINER, J., NOVOA BARSOTTINI, C. G., LACERDA, D., and MAGALHÃES de MARCO, L. R. (2009). Empirical Evaluation in Computer Science Research Published by ACM. Information and Software Technology. V.51, n.6, p.1081-1085, 2009.

WAINER, J., XAVIER, E. C., BEZERRA, F. Scientific Production in Computer Science: A Comparative Study Between Brazil and Other Countries. Scientometrics, 2009.

WEKA. Waikato Environment for Knowledge Analysis. Disponível em: <www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 5 ago. 2009.