XVII COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA
Universidade, desenvolvimento e futuro na Sociedade do Conhecimento
Mar del Plata – Argentina
22, 23 e 24 de novembro de 2017
ISBN: 978-85-68618-03-5

SEGMENTAÇÃO DE DADOS PARA A PREDIÇÃO DE FUTURAS PUBLICAÇÕES POR PARTE DOCENTE DA INSTITUIÇÃO

GUILHERME AUGUSTO BAUER
Universidade de Santa Cruz do Sul
guiabauer@gmail.com

EDUARDO KROTH
Universidade de Santa Cruz do Sul
kroth@unisc.br

JACQUES NELSON CORLETA SCHREIBER
Universidade de Santa Cruz do Sul
jacques@unisc.br

ELPIDIO OSCAR BENITEZ NARA
Universidade de Santa Cruz do Sul
elpidio@unisc.br

ALVIN LAURO BESKOW
Universidade de Santa Cruz do Sul
alvinbeskow@hotmail.com

JOÃO VICTOR KOTHE
Universidade de Santa Cruz do Sul
joaokothe@mx2.unisc.br

RESUMO
A necessidade de as organizações terem informação correta e de forma rápida se repete nas universidades. A prática demonstra que existe uma dificuldade em se conseguir informações referentes a produção bibliográfica, científica e tecnológica de seus professores. Esses dados podem ser encontrados na Plataforma Lattes que possui na sua base de dados, os currículos de pesquisadores e professores brasileiros. Para realizar a extração de conhecimento de uma grande base de dados podem ser utilizadas técnicas de KDD (Descoberta de Conhecimento em Bases de Dados). Este artigo apresenta a sugestão de uma solução em software para coletar dados dos currículos Lattes dos professores de uma Instituição de Ensino Superior brasileira. O software implementa algoritmos de mineração de dados sobre os dados coletados a fim de gerar conhecimento que poderá ser utilizado em tomadas de decisão por gestores da universidade. Este artigo tem como principal objetivo, apresentar um método para obtenção de padrões de sequência de publicações qualificadas a partir de dados reais coletados na Plataforma Lattes, para aprimorar a gestão acadêmica dos cursos de Stricto Sensu (mestrado e doutorado). A tese inspiradora foi que, conhecendo-se o *timeline* de publicações do docente possa-se antever quais publicações qualificadas serão produzidas num futuro próximo.

Palavras chave: Sequenciação, KDD, Lattes, Mineração, Docentes.

1. INTRODUÇÃO

Devido à grande competitividade existente no atual mercado, possuir informações corretas e de forma rápida para realizar tomadas de decisões é de suma importância para as organizações. Entre essas organizações estão as universidades que entre outros dados desejam ter conhecimento da produção científica, bibliográfica e tecnológica de seu corpo docente. E com o grande crescimento no volume de dados que circulam pelas organizações e o consequente aumento no tamanho de suas bases de dados é necessário o uso de tecnologias de informação para conseguir coletar, gerir, analisar, armazenar e utilizar da melhor forma possível essas informações; é nesse contexto que surgem ferramentas para auxílio nesses diversos processos.

Uma dessas possibilidades que podem ser utilizadas é o KDD (*Knowledge Discovery in Databases*), processo utilizado para descobrir conhecimento em grandes bases de dados. Entre as etapas do processo de KDD a principal delas é o *Data Mining* ou Mineração de Dados que consiste no uso de algoritmos específicos para localizar padrões de comportamento nos dados disponíveis. Dentre as técnicas da Mineração de Dados foram selecionadas para uso nesse artigo a Descoberta de Associações e a Sequenciação.

Muitos dos dados referentes à produção científica, bibliográfica e tecnológica do corpo docente de uma universidade podem ser encontrados nos currículos dos professores cadastrados na Plataforma Lattes, que é uma ferramenta desenvolvida pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e que possui um módulo direcionado ao cadastro de currículos de pesquisadores brasileiros com dados referentes a sua produção científica, bibliográfica e tecnológica.

O artigo é focado em um estudo das técnicas de KDD e Mineração de Dados tendo como objetivo principal a aplicação de algoritmos de *Data Mining* a fim de identificar conhecimento útil a partir de dados de currículos Lattes de professores de uma IES brasileira, com o principal objetivo de auxiliar o gestor, em saber qual será a produção acadêmica dos docentes, para que com essas informações ele possa gerenciar de forma proativa as suas metas.

2. FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica está dividida em duas etapas, a descrição da técnica de Descoberta de Conhecimento em Bases de Dados e posteriormente e a descrição do Currículo Lattes.

2.1 KDD (Knowledge Discovery in Databases)

O desenvolvimento desse projeto foi baseado na metodologia do KDD, onde é dividido em 5 principais etapas.

2.1.1. Seleção de Dados

Onde tem-se os dados da organização de forma bruta, mas precisamos identificar quais são as reais informações que implicam na solução do problema, removendo campos com uma correlação baixa com a solução da proposta, e adicionar informações que tem uma correlação relevante com o problema.

2.1.2. Processamento e Limpeza de Dados

Nessa etapa já possuímos algumas informações relacionadas a quais campos são mais relevantes, no entanto, esses dados podem haver informações problemáticas, como um valor fora do permitido, e por isso precisamos corrigir essas informações. Segundo Navega (2002, apud Costa et al.) "as bases de dados são dinâmicas, incompletas, redundantes, ruidosas e esparsas, necessitando de um pré-processamento para limpá-las"

2.1.3. Transformação de Dados

Os campos mais relevantes para a solução do problema já foram selecionados e suas informações corrigidas para ficarem dentro do limite permitido, então, é importante transformar os dados em informações legíveis para o algoritmo de mineração que será usado, isso pode ser feito através da discretização, normalização ou qualquer outra técnica de preparação de dados. Segundo Camilo & Silva (2009), algumas das técnicas que podem ser usadas nesse passo são: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores genéricos), normalização (coloca as variáveis em uma mesma escala), e criação de novos atributos a partir de atributos já existentes.

2.1.4. Data Mining

A partir dos dados, usa-se algoritmos e algumas técnicas de mineração de dados para a refinação e, com isso, obtém-se um conhecimento implícito nas bases de dados.

A tese inspiradora deste projeto de pesquisa foi: "é possível prever quais poderão ser as futuras publicações dos docentes, assim como se prevê comportamento futuro dos consumidores? " Buscando provar essa tese, o presente trabalho foi baseado em um algoritmo de sequenciação denominado *Generalized Sequential Patterns* (GSP), que é uma ramificação do algoritmo Apriori. De acordo com Barbosa (2006), o algoritmo GSP foi uma das primeiras estratégias utilizadas para descoberta e extração de padrões sequenciais.

O algoritmo GSP foi baseado no algoritmo Apriori, com isso, herdou algumas características, como por exemplo o cálculo de Suporte que é muito usado para identificar a relevância de uma informação perante a base de dados. O cálculo do suporte, pode ser definido simplesmente, pela soma das aparições de cada item por candidato, ou pela proporção da aparição do mesmo, para então, determinar-se um valor mínimo de corte.

O processo de mineração de dados por sequenciação requer que o desenvolvedor, crie uma linha temporal, determinando os itens, e segmentos temporais de cada candidato. Por exemplo, como determinar quais itens o cliente 1 irá comprar, caso já tenha comprado uma barra de chocolate na primeira compra, e na segunda compra, uma barra de chocolate e um pote de leite condensado? Simples, cria-se um grupo de clientes com um perfil semelhante, e é determinado um limite temporal, para tais ocorrências. Por exemplo: *<a(ab)(ac)bc>*, onde *"a"* foi a ocorrência de compra de chocolate, em seguida a ocorrência *"ab"*, e pode-se notar que ela é composta por dois itens, depois *"ac"*, *"b"*, e por fim *"c"*. Ou seja, nessa abordagem cada letra é equivalente a um item, e todo o conteúdo interno dos parênteses, equivalem a uma compra, ou a um limite temporal, que une as informações de compra. Quando não há parênteses, um item também simboliza um limite temporal.

O processo do GSP é divido em dois principais passos; O primeiro, no qual consiste em gerar os candidatos, e o segundo no qual consiste em podar os candidatos menos relevantes, em relação ao suporte mínimo. A figura 1, ilustra bem as duas principais etapas. A primeira coluna refere-se à base de dados, onde cada linha é um candidato, e está delimitado

por itens, e por laços temporais, na segunda coluna os candidatos gerados a partir da base, e pôr fim a poda desses candidatos, com relação ao Suporte mínimo necessário.

Frequent 3-Sequences	Candidate 4-Sequences	
	after Join	after pruning
<(1,2)(3)>	<(1,2)(3,4)>	<(1,2)(3,4)>
<(1,2)(4)>	<(1,2)(3)(5)>	
<(1)(3,4)>		
<(1,3)(5)>		
<(2)(3,4)>		
<(2)(3)(5)>		

Figura 1: Processo GSP
Fonte: Srikant, R., & Agrawal, R. (1996)

Passo 1 – Geração dos candidatos

Nessa etapa, é iniciado o processo de mineração com os dados da primeira coluna, com isso, considera-se que as linhas da base sejam conjuntos $W1$ e $W2$, e cada item dentro das ocorrências dos conjuntos sejam subconjuntos de $W1$, t. Então, temos $W1 = \{t_0, t_1, \dots, t_n\}$, onde, do exemplo anterior, t_1 equivale a "a". Com isso, unem-se as palavras dos conjuntos, iniciando o conjunto $W1$ em $n = 1$, mas além disso, o conjunto $W2$ termina em t_{n-1}. Ou seja, iremos comparar $t_1, t_2, \dots, t_n$ do conjunto $W1$, com $t_0, t_1, \dots, t_{n-1}$ do conjunto $W2$. Caso sejam iguais os valores das ocorrências e itens, temos a junção das informações $W1$ e $W2$, formando um novo candidato. Essa etapa será feita para todas as combinações, e irá gerar a segunda coluna, a coluna de geração candidatos da figura 1.

Passo 2 – Poda de candidatos

Nessa etapa analisamos os dados gerados a partir da etapa anterior, para determinar se esse candidato realmente é relevante. Para isso é preciso identificar quantos itens existem em cada candidato, supondo que seja K, então precisa-se transformar esse novo candidato em uma matriz KxK, onde cada linha receberá o conteúdo da primeira, então haverá uma "redundância" de informação, no entanto, após isso estar formado, remove-se a diagonal principal da matriz, transformando a antiga matriz KxK para uma Kx(K-1), essa nova matriz deverá ter seus itens (colunas) novamente agrupados, com isso haverá novamente K candidatos mas todos com n-1 itens.

Após a transformação dessa matriz, analisa-se cada novo candidato com a base de dados, e verificamos quantos deles existem na base, para então ter-se o suporte de cada novo candidato gerado, para então podá-lo ou mantê-lo. No próximo tópico será abordado a parte de análise dos dados**.**

2.1.5. Interpretação

Segundo Nicolaio, Pelinski (2006, apud Macedo, Matos, 2010), nessa etapa é a fase onde é realizada a interpretação final dos padrões de conhecimento coletados e a verificação da necessidade de repetir os passos de processo de mineração para tentar adquirir conhecimento mais útil e de forma mais clara para o usuário ou a finalização do processo para que o conhecimento possa ser utilizado como apoio na tomada de decisão.

2.2. PLATAFORMA LATTES

A Plataforma Lattes integra bases de dados de currículos de pesquisadores, de grupos de pesquisa e de instituições em um único sistema de informação desenvolvido pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Essa plataforma foi lançada em agosto de 1999 com a primeira versão do currículo Lattes. A plataforma é composta pela integração de quatro sistemas distintos: Currículo Lattes; Diretório de grupos de pesquisa; Diretório de instituições; Sistemas gerencial de fomento.

A figura 2 apresenta um esquema dos dados existentes no currículo Lattes, suas relações e hierarquia.

Figura 2: Esquema das relações dos dados no currículo Lattes
Fonte: Autores, 2017.

2.3. QUALIS

Qualis é um sistema criado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e que é utilizado para classificar a produção científica de programas de pós-graduação com base nos artigos publicados em periódicos científicos.

A classificação é realizada por comitês de consultores de cada área de avaliação de acordo com critérios definidos pela área e aprovados pelo Conselho Técnico-Científico da Educação Superior (CTC-ES) que procuram refletir a importância relativa dos diferentes periódicos para uma determinada área.

A estratificação da qualidade dessa produção é realizada de forma indireta. Dessa forma, o Qualis afere a qualidade dos artigos e de outros tipos de produção, a partir da análise da qualidade dos veículos de divulgação, ou seja, periódicos científicos.

A classificação de periódicos é realizada pelas áreas de avaliação e passa por processo anual de atualização. Esses veículos são enquadrados em estratos indicativos da qualidade - A1, o mais elevado; A2; B1; B2; B3; B4; B5; C - com peso zero. E o mesmo periódico pode ser classificado em duas ou mais áreas diferentes e receber diferentes avaliações em cada uma dessas áreas.

A função do QUALIS é exclusivamente para avaliar a produção científica dos programas de pós-graduação no Brasil.

3. METODOLOGIA

Foi realizada uma pesquisa quantitativa e exploratória com pesquisa bibliográfica referente aos assuntos tratados neste artigo e em um momento posterior com relação a geração de conhecimento útil através dos dados de uma base de dados utilizando técnicas de KDD (Descoberta de Conhecimento em Bases de Dados) e *Data Mining*.

Inicialmente foi realizado um levantamento bibliográfico em livros, artigos e outras publicações referentes aos seguintes assuntos que são abordados neste artigo: KDD, Mineração de Dados com foco em Clusterização e Regras de Associação, Sequenciação e, a Plataforma Lattes focando nos currículos de professores e pesquisadores. Após esse levantamento foram analisadas as referências coletados selecionando as que realmente tiveram utilidade para o desenvolvimento do artigo e foi realizada a fase de escrita do mesmo.

Posteriormente foram definidas a forma de como será feita a coleta dos dados necessários para a geração do conhecimento útil, as técnicas que foram utilizadas nesse processo e ocorreu o desenvolvimento de uma ferramenta para extração dos dados dos currículos Lattes de professores e pesquisadores, e para a aplicação, sobre esses dados, de algoritmos de clusterização e de regras de associação para possibilitar a geração de conhecimento útil a partir dos dados que é o objetivo principal deste artigo.

3.1. VISÃO GERAL

A ferramenta desenvolvida neste artigo realiza a extração de dados referentes à produção científica, bibliográfica e tecnológica dos currículos Lattes dos professores. Os dados são lidos de arquivos XMLs dos currículos Lattes, estruturados e salvos em uma base de dados modelada e construída anteriormente. A ferramenta também permite que sobre os dados extraídos sejam aplicados algoritmos de *Data Mining*. A figura 3 mostra um esquema básico de sequência das operações que são realizadas pelo sistema desenvolvido.

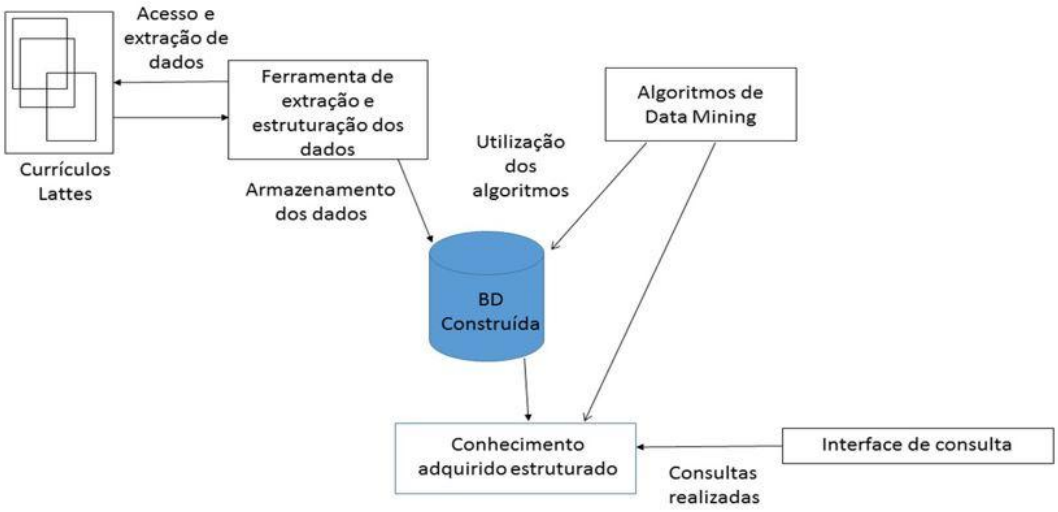

Figura 3: Esquemas dos passos da solução proposta
Fonte: Autores, 2017.

Os algoritmos a serem executados e os dados sobre os quais eles serão aplicados poderão ser definidos pelo usuário a partir de opções predefinidas na criação da ferramenta. O conhecimento gerado com a aplicação dos algoritmos será estruturado e exibido ao usuário de forma textual e/ou gráfica.

3.2. FUNCIONALIDADE

O sistema desenvolvido permite o download dos currículos Lattes dos professores através de seus CPFs ou de seus nomes. Na interface destinada a este fim podem ser informados os dados de diversos professores para a realização do download de todos os currículos em um mesmo processo e também pode ser definido o local onde os arquivos dos currículos deverão ser salvos.

3.2.1. Processo de extração e carga dos dados na base de dados

No sistema desenvolvido existe a opção de carregar arquivos XMLs dos currículos para serem lidos, estruturados e salvos na base de dados previamente construída. Na interface desenvolvida para este processo existe a opção de selecionar o local onde estão salvos os arquivos, será realizada uma verificação no formato e estrutura dos arquivos para verificar quais são os arquivos dos currículos e os dados destes arquivos são carregados em uma tabela onde existe a possibilidade de selecionar os arquivos que serão lidos e estruturados para serem salvos na base de dados. Para facilitar a identificação dos arquivos uma das informações carregadas nesta tabela de seleção é o nome dos professores. Nesta mesma interface existe uma tabela de consulta na qual estarão disponíveis os últimos currículos importados ou atualizados com o nome do professor e a data da última atualização.

A verificação da estrutura e a leitura dos arquivos são realizadas usando como base um arquivo XML *Schema Definition* (XSD) que é um arquivo que define o formato padrão que um arquivo XML deve seguir. Nesse arquivo são indicados os nodos principais do arquivo XML na *tag element*, os subnodos e atributos destes nodos principais indicando, quando necessário, o número mínimo e máximo de ocorrências do elemento ou atributo. Também

podem existir neste arquivo de esquema o tipo de dado dos atributos e se a existência desses atributos é obrigatória ou opcional no arquivo XML que deve seguir o esquema.

O arquivo XSD também foi utilizado para geração automática das classes Java utilizadas inicialmente para a leitura dos dados e para que eles possam ser manipulados nas demais classes do sistema.

3.2.2. Aplicação de algoritmos da Data Mining

Existe no sistema a possibilidade de utilização de algoritmos de *Data Mining* sobre os dados salvos. O algoritmo disponibilizado, para aplicação sobre os dados, é o algoritmo de descoberta de sequências temporais *Generalized Sequential Patterns* (GSP), que pode ser aplicado sobre os dados das produções científicas, tecnológicas e bibliográficas dos professores para que através da classificação Qualis de suas produções atuais seja possível prever as prováveis classificações de suas publicações futuras. Neste artigo estão sendo utilizadas as classificações Qualis definidas pela CAPES (A1, A2, B1, B2, B3, B4, B5, C) e a classificação Não Classificado (NC) para as produções que não possuem classificação Qualis, transformadas para números inteiros de 1 a 9 que são aceitos para a execução do algoritmo para prever as classificações das produções futuras.

Os dados da classificação Qualis são pesquisados a partir do código *International Standard Serial Number* (ISSN) dos periódicos onde foram publicadas as produções dos professores, usando como base a última versão da classificação existente, Qualis 2014.

O sistema também possui pronta a base para a implementação e uso de algoritmos de clusterização e descoberta de associações como por exemplo o algoritmo Apriori.

Para a aplicação dos algoritmos está sendo utilizada a ferramenta WEKA, desenvolvida pela Universidade de Waikato, na Nova Zelândia e que possui a implementação de diversos algoritmos de técnicas de Data Mining, principalmente para as técnicas de Classificação, Clusterização e Descoberta de Associações.

3.3. FERRAMENTAS E SOFTWARES UTILIZADOS

O sistema foi totalmente desenvolvido na linguagem de programação Java e foi utilizado o *framework Hibernate* para realizar o mapeamento entre as classes Java e as tabelas da base de dados com o uso de arquivos XML do tipo *Hibernate Mapping* (HBM) e as consultas aos dados estão sendo realizadas sobre as classes Java com o uso da linguagem *Hibernate Query Language* (HQL) permitindo portabilidade de banco de dados.

Para a construção da base de dados e armazenamento dos dados extraídos dos currículos Lattes foi utilizado um banco de dados relacional Sistema de Gerenciamento de Banco de Dados (SGBD) MySql.

3.4. TESTES E VALIDAÇÕES

Inicialmente foram realizados testes com pequenos grupos de professores para verificar o correto funcionamento do sistema e do algoritmo de mineração. Para as validações do sistema e do algoritmo foram realizadas duas fases de testes na extração dos dados dos currículos e aplicação do algoritmo GSP.

Na primeira dessas fases foram utilizados os dados dos currículos dos onze professores de um programa de mestrado da IES utilizada na validação do sistema elaborado.

Na segunda fase foram realizados testes com uma quantidade maior de dados, para esses testes foram utilizados os dados de quarenta e dois professores do Programa de Pós-Graduação em Computação (PPGC) de outra IES brasileira.

Os dados da classificação Qualis são coletados usando como base a última versão da classificação existente, Qualis 2014 através do código ISSN dos periódicos. Esses dados são adicionados manualmente na base de dados.

Para realizar a previsão o algoritmo trabalha com dados numéricos e por esse motivo as classificações Qualis definidas pela CAPES (A1, A2, B1, B2, B3, B4, B5, C) e a classificação Não Classificado usada para as produções que não possuem classificação Qualis, são transformadas para números inteiros de 1 a 9 que são aceitos para a execução do algoritmo.

Para a correta execução do algoritmo é necessário que os registros possuam data e hora completa, como as produções no currículo Lattes possuem apenas o ano de publicação inicialmente é realizado um processamento sobre os dados para que a variável de data possua um formato aceito pelo algoritmo.

4. RESULTADOS

Nos testes realizados o nível de confiança dos resultados foi definido como 90% e os resultados destes testes com a previsão do nível da classificação Qualis das produções futuras dos professores da IES estão na tabela 1 e dos professores da outra instituição na tabela 2.

Professor	**Qualis Últimas Produções**	**Qualis Produções Futuras**
Professor 1	NC, A2, A2, A2, B3, B5	B3, B2, B2
Professor 2	B4, A1, B1, B3	B2, B2, B2
Professor 3	NC, NC, NC, NC, NC, NC	NC, B5, B4
Professor 4	B5, B5, B2, C, C, C	B4, B4, B4
Professor 5	C, B2, B4, C, NC, NC	B5, B5, B5
Professor 6	A1, B3, B4, C, NC, NC	B4, B4, B4
Professor 7	B4, B5, B5, C, NC, NC	B5, B4, B4
Professor 8	NC, B4, B4, NC, B5, C	B5, B5, C
Professor 9	B5, B5, C, NC, B5, NC	B5, B5, B5
Professor 10	C, NC, B1, B5, NC, NC	B5, B5, B5
Professor 11	B5, C, A2, B5, B5, A1	B3, B4, B5

Tabela 1: Resultados dos Testes IES 1
Fonte: Autores, 2017.

Professor	Qualis Últimas Produções	Qualis Produções Futuras
Professor 1	NC, A1, A2, A1, B2, A2	B1, B1, B1
Professor 2	C, A1, NC, NC, C, C	B5, C, B5
Professor 3	NC, NC, A1, A2, NC, B1	B5, C, A1
Professor 4	B2, NC, B2, A2, NC, A1	B4, C, A2
Professor 5	B4, A2, NC, B4, A2, A2	B4, B4, A2
Professor 6	B2, B1, B1, NC, NC, B1	B1, B1, B1
Professor 7	NC, B1, A2, B1, A2, A1	B3, A2, B3
Professor 8	NC, B4, B2, B5, NC, A2	B3, B3, B4
Professor 9	A2, C, A1, A1, NC, NC	B3, NC, B4
Professor 10	B5, B4, B1, C, B1, NC	B1, NC, C
Professor 11	C, A2, B4, C, B1, B2	C, B1, C
Professor 12	B1, A2, B1, A2, A1, NC	B3, B2, B3
Professor 13	C, B1, A2, B1, NC, A2	B1, B2, B4
Professor 14	B1, A2, B1, B1, B1, B1	A2, B1, B1
Professor 15	NC, B5, B1, A2, B2, NC	B4, B2, B4
Professor 16	B3, A2, B1, A2, B2, B2	B1, B1, B1
Professor 17	A2, A2, NC, B1, B1, NC	B1, B1, A2
Professor 18	NC, A2, B1, A1, A1, NC	B3, B2, A2
Professor 19	NC, NC, C, B5, NC, B1	B5, NC, B3
Professor 20	A2, A2, B1, B1, A2, A1	B3, A2, B1
Professor 21	NC, A2, B1, A1, A2, A1	A2, A1, A2
Professor 22	A2, B2, B2, NC, B2, B1	B3, B2, B3
Professor 23	NC, C, A2, A2, NC, NC	B4, B4, B4
Professor 24	B1, B1, A2, NC, C, A1	B4, B4, B5
Professor 25	A2, A2, B2, A1, B3, A2	B1, B1, A2
Professor 26	B1, B1, C, B1, A1, B2	B1, B1, A2
Professor 27	NC, B1, NC, NC, B2, A2	B4, NC, B4
Professor 28	A2, A2, B1, A2, A1, A2	A2, A2, B1
Professor 29	A2, A2, B1, A2, A2, A2	B4, A2, A1
Professor 30	B5, A2, B5, A2, B1, B5	B4, B4, B5
Professor 31	A2, A2, A2, B3, B3, B1	B2, B1, B1

Professor 32	B1, NC, A2, B3, B1, A1	B2, B2, B2
Professor 33	B1, B1, B2, A1, B1, A1	A1, A1, B1
Professor 34	B1, A2, B2, NC, NC, NC	A2, B3, B4
Professor 35	B4, NC, B1, C, C, A1	C, B1, B4
Professor 36	B1, NC, B1, A2, NC, A1	B3, B3, B1
Professor 37	A2, NC, B2, A2, A2, A2	B1, B1, B2
Professor 38	NC, NC, B1, C, A2, A2	B4, B4, B2
Professor 39	B2, NC, A2, B1, A2, A2	B2, B2, B2
Professor 40	A2, B1, B1, NC, NC, B1	A1, B4, NC
Professor 41	B2, B1, B1, B1, B5, B3	B2, B1, B2
Professor 42	NC, A2, B1, B2, B2, C	B2, B3, B2

Tabela 2: Resultados dos Testes IES 2
Fonte: Autores, 2017.

Os resultados exibidos pela execução do algoritmo não são exatos e foram arredondados para definir a classificação Qualis das produções futuras. O resultado do algoritmo é gerado da forma como está demonstrado na figura 4. Nesta imagem estão as datas, já transformadas para o formato aceito, das produções de um professor e a classificação Qualis, também já transformada para o formato aceito, destas produções. E nas últimas três linhas estão os resultados de previsão das classificações das próximas produções do professor.

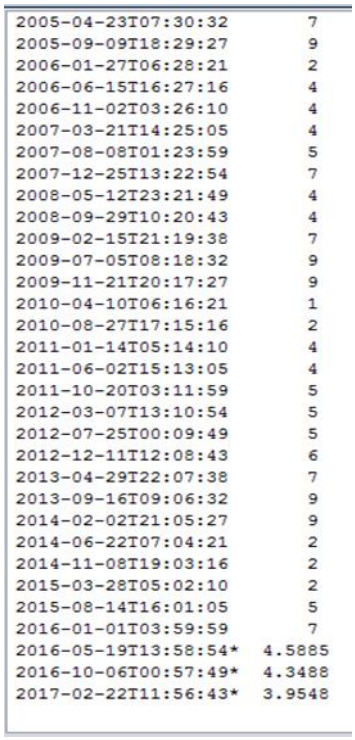

Figura 4: Resultados da execução do algoritmo
Fonte: Autores, 2017.

5. CONCLUSÃO

Atualmente existe uma dificuldade das universidades de possuir um controle referente à produção bibliográfica, científica e tecnológica de seus professores e de ter a possibilidade de utilizar esses dados para auxiliar em alguma tomada de decisão, devido à grande quantidade de departamentos e cursos existentes e aos mais diversos eventos e publicações em que os professores podem apresentar e publicar seus trabalhos e artigos.

Muitos desses dados podem ser encontrados nos currículos Lattes dos professores existentes na Plataforma Lattes, base de dados de currículos e grupos de pesquisa mantida pelo CNPq e que possui diversos dados referentes à produção bibliográfica, científica e tecnológica de professores e pesquisadores.

Para a extração, estruturação e processamento desses dados podem ser utilizadas técnicas de Descoberta de Conhecimento em Bases de Dados, principalmente Mineração de Dados que é o principal tópico estudado neste artigo com foco em algoritmos de Clusterização, para agrupamento de dados similares em bases de dados, e de Regras de Associação para localizar relações entre ocorrências simultâneas de conjuntos de dados em transações de bancos de dados.

Neste artigo foi desenvolvido uma solução para essa dificuldade das universidades controlarem a produção dos professores por meio de uma ferramenta que realizará a extração dos dados dos currículos Lattes dos professores, realizando o armazenamento dos dados coletados em uma base de dados e posteriormente aplicar sobre os dados algoritmos de Clusterização e Regras de Associação para extrair conhecimento útil para tomadas de decisão que poderão ser estruturados e visualizados através de consultas realizadas pelos usuários em uma interface própria para esse fim.

REFERÊNCIAS

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

MACEDO, Dayana Carla; MATOS, Simone Nasser. Extração de conhecimento através da mineração de dados. Revista de Engenharia e Tecnologia, v. 2, n. 2, p. Páginas 22-30, 2010.

NAVEGA, Sergio. Princípios essenciais do data mining. Anais do Infoimagem, 2002.

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh. Mining sequential patterns: Generalizations and performance improvements. Advances in Database Technology—EDBT'96, p. 1-17, 1996.