PVAF Manager - Um Sistema de Gerenciamento de Informação sobre Veículos de Publicação Científica

Alternative Title: PVAF Manager - An Information Management System on Scientific Publication Venues

Tiago Antônio Paraizo
Depto Ciência da Computação
Univ. Federal de Lavras
Lavras, MG, Brazil
paraizocomp@gmail.com

Douglas Henrique Silva
Depto Ciência da Computação
Univ. Federal de Lavras
Lavras, MG, Brazil
douglas13mg@gmail.com

Denilson Alves Pereira
Depto Ciência da Computação
Univ. Federal de Lavras
Lavras, MG, Brazil
denilsonpereira@dcc.ufla.br

RESUMO

Um arquivo de autoridade de veículos de publicação armazena variações nos nomes de periódicos e conferências que publicam artigos científicos. É útil na construção de ferramentas de busca e desambiguação de dados, e é de especial interesse de agências de fomento à pesquisa e de avaliação de programas de pós-graduação, as quais usam a qualidade dos veículos de publicação como base para avaliação de publicações de pesquisadores e grupos de pesquisa. Entretanto, manter um arquivo de autoridade atualizado não é uma tarefa trivial. Diferentes nomes são usados para se referenciar um mesmo veículos de publicação, algumas vezes eles mudam de nome, novos veículos surgem regularmente e os seus índices de qualidade são atualizados frequentemente. Este trabalho apresenta o desenvolvimento do PVAF Manager, um sistema de gerenciamento de informações sobre veículos de publicação científica. Ele representa a evolução de um trabalho anterior, e incorpora características como a expansão do banco de dados de veículos de publicação para todas as áreas do conhecimento cobertas pelo Qualis Capes, ferramentas para atualização anual de índices bibliométricos, opção para atualização e correção de dados, gerenciamento e coleta de sugestões de usuários. São apresentados os detalhes de implementação e uma avaliação experimental dos resultados da expansão e da qualidade do método de consulta aos dados. Os resultados mostram uma boa cobertura do PVAF em relação a veículos de publicação internacionais e baixas taxas de erro do método de consulta aos dados.

Palavras-Chave

arquivo de autoridade, veículos de publicação, desambiguação de dados, Qualis Capes

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

ABSTRACT

A publication venue authority file stores variations in the names of journals and conferences that publish scientific articles. It is useful in the construction of search tools and data disambiguation, and it is of special interest to agencies funding research and evaluating graduate programs, which use the quality of publication venues as a basis for evaluation of publications of researchers and research groups. However, keeping an updated authority file is not a trivial task. Different names are used to reference a same publication vehicles, sometimes they change their name, new venues emerge regularly and their quality indexes are updated frequently. This paper presents the development of PVAF Manager, a system for managing information about scientific publication venues. It represents the evolution of a previous work, and incorporates features such as the expansion of the publication venue database for all areas of knowledge covered by Qualis Capes, tools for annual updating of bibliometric, option for updating and correcting data, management and collecting of user suggestions. We present the implementation details and an experimental evaluation of the results of the expansion and the quality of the data search method. The results show a good coverage of PVAF in relation to international publication venues and low error rates of the data search method.

CCS Concepts

•**Information systems** → **Information integration;** *Digital libraries and archives; Retrieval tasks and goals;*

Keywords

authority file, publication venue, data disambiguation, Qualis Capes

1. INTRODUÇÃO

Agências de fomento à pesquisa e de avaliação de programas usam o impacto de publicações científicas para tomarem decisões sobre financiamentos de projetos e para avaliarem programas de pós-graduação. No Brasil, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes)[1]

[1]http://www.capes.gov.br/

criou o Qualis, um sistema de indexação usado para classificar a produção científica dos programas de pós-graduação com base na reputação do veículo de publicação dos artigos publicados por cada programa. Outros índices reconhecidos internacionalmente também avaliam a qualidade de veículos de publicação, tais como o Journal Impact Factor (JIF)[2], o Scopus CiteScore[3], o SCImago Journal & Country Rank (SJR)[4] e o H5 Google Scholar[5].

Entretanto, para que esses índices possam ser usados de forma eficaz por sistemas de recuperação de informação, é necessário reconhecer o nome do veículo de publicação (periódico, conferência, workshop) na cadeia de caracteres referente a cada citação bibliográfica. Erros de extração automática, abreviações, uso de siglas e as diversas formas de se escrever uma referência para um mesmo veículo de publicação contribuem para tornar essa tarefa mais difícil. Por exemplo, "Simpósio Brasileiro de Sistemas de Informação", "SBSI", "Brazilian Symposium on Information Systems", "Simpósio de Sistemas de Informação (SBSI)" e "Simp. Bras. Sist. Inf." são referências distintas para um mesmo veículo de publicação.

Uma forma encontrada por bibliotecas digitais para lidar com referências distintas para uma mesma entidade é usar arquivos de autoridade. Um arquivo de autoridade mantém uma lista das variações de escrita usadas para um atributo bibliográfico específico (ex: nome de autor, veículo de publicação ou afiliação) [2]. O VIAF [22] é um exemplo de um esforço internacional para integração de arquivos de autoridade, o qual mantém variações de nomes de autores.

Em [18], os autores apresentam o PVAF, um arquivo de autoridade de veículos de publicação para a área de Ciência da Computação. O PVAF consiste de um conjunto de registros com informações sobre variações de nomes e siglas usadas correntemente, nomes e siglas antigos, índices bibliométricos, dentre outras informações sobre veículos de publicação. A Figura 1 ilustra um exemplo de um registro nesse arquivo de autoridade, obtido pela consulta "Logic Journal of the IGPL" na ferramenta disponível em http://pvaf.dcc.ufla.br. Entretanto, o arquivo de autoridade criado por [18] está restrito a veículos de publicação da área de Ciência da Computação, não possui ferramentas para atualização anual dos índices bibliométricos e nem para correções de dados.

Este trabalho apresenta o PVAF Manager, um sistema de gerenciamento de informação de veículos de publicação científica. Ele é uma evolução do trabalho de [18], e incorpora as seguintes características: (i) expansão do banco de dados de veículos de publicação cobrindo outras áreas do conhecimento, por meio da importação dos veículos de publicação indexados no Qualis Capes; (ii) ferramentas para a atualização anual dos índices bibliométricos; (iii) opção para atualização e correção de dados; (iv) gerenciamento de usuários aptos a atualizar os dados e (v) coleta de sugestões de usuários.

Na importação e atualização dos dados, é necessário tratar a desambiguação de nomes. Os veículos de publicação no arquivo de autoridade não possuem um atributo identificador (chave). Embora os periódicos tenham o ISSN, o qual possui

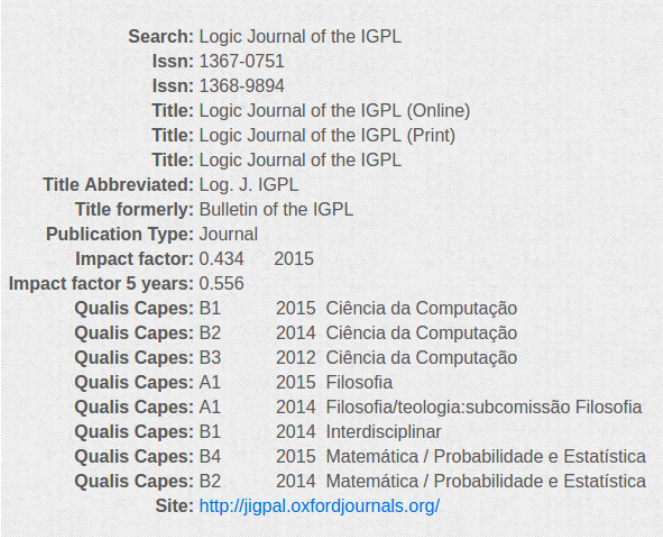

Figura 1: Exemplo de um registro no PVAF, obtido pela consulta "Logic Journal of the IGPL".

um valor único para cada periódico, existem casos em que o ISSN não é conhecido, casos de periódicos que possuem mais de um ISSN, para as versões impressa e online, e casos de periódicos que mudaram de ISSN. Para as conferências e workshops, é ainda mais complexo, pois os nomes e siglas não são únicos. Embora o Qualis adote a sigla como identificador, existem conferências distintas com a mesma sigla e existem conferências com mais de uma sigla, as quais são adotadas por diferentes bibliotecas digitais ou devido a mudanças durante sua história. Veículos de publicação, muitas vezes, mudam de nome, duas ou mais conferências são juntadas para formar uma outra, ou são separadas, formando duas ou mais outras novas, e workshops são transformados em conferências. Assim, manter a consistência dos dados não é uma tarefa trivial. Neste trabalho, são apresentadas as estratégias usadas para lidar com essas questões.

A manutenção de um arquivo de autoridade de veículos de publicação se justifica devido a sua importância como fonte de dados para sistemas de busca e recuperação de informação bibliográfica e para elaboração de ferramentas de desambiguação de dados. Pode ser usado, por exemplo, para criação de uma ferramenta de comparação de grupos de pesquisa com base na qualidade de suas publicações, o que seria muito útil na avaliação dos programas de pós-graduação no Brasil.

Em suma, as principais contribuições deste trabalho são:

- o desenvolvimento de um conjunto de métodos e software para expansão do arquivo de autoridade criado por [18], de forma a cobrir novas áreas do conhecimento e para atualização constante de seus dados;
- a avaliação experimental dos resultados da expansão e da qualidade do método de consulta aos dados;
- a disponibilização do PVAF expandido e de suas novas funcionalidades para a comunidade científica, a qual poderá consultar os dados e sugerir correções e atualizações nos mesmos.

[2]http://wokinfo.com/products_tools/analytical/jcr
[3]https://journalmetrics.scopus.com
[4]http://www.scimagojr.com
[5]https://scholar.google.com/intl/en/scholar/metrics.html

2. TRABALHOS RELACIONADOS

Bibliotecas digitais coletam e armazenam dados sobre publicações científicas, e precisam manter a sua consistência. Em [17], os autores destacam os principais problemas envolvidos nessa tarefa. Dentre eles destacam-se os erros na entrada de dados, falta de padrões, diversos estilos de citação, software coletores e extratores imperfeitos, nomes de autores ambíguos e abreviações de nomes de veículos de publicação.

A adoção de arquivos de autoridade [2] é uma solução usada por algumas bibliotecas digitais para manter a consistência dos dados. Eles armazenam uma lista de variações de rótulos usadas para referenciar um item bibliográfico. O Virtual International Authority File (VIAF) [4, 22] é um exemplo, o qual integra arquivos de autoridade de bibliotecas nacionais de vários países, criando um serviço de autoridade único, disponível na Web. O foco do VIAF é em nomes de autores, o que o difere do arquivo de autoridade discutido neste trabalho, cujo foco é em veículos de publicação científica.

Arquivos de autoridade contribuem na vinculação e compartilhamento de dados na Web, como mostra [11]. No trabalho, os autores apresentam um estudo sobre aplicações de *Linked Data* [5] em bibliotecas digitais. *Linked Data* permitem às bibliotecas digitais resolver consultas mais complexas por meio de conexões com fontes de dados externas. Eles mostraram que na maioria das implementações faltam ferramentas de suporte e mecanismos de controle de qualidade dos dados. Além disso, as aplicações não são capazes de receber colaborações de usuários, e que essa interação pode ser proveitosa para o enriquecimento dos dados. O PVAF Manager, implementado neste trabalho, procura tratar essas questões e, portanto, pode contribuir para melhorar as aplicações de bibliotecas digitais.

Este trabalho vem de uma sequência de outros trabalhos envolvendo arquivos de autoridade de veículos de publicação. Em [19], os autores apresentam um método para criar arquivos de autoridade de veículos de publicação a partir dos dados de um conjunto de citações bibliográficas e de informações extraídas da Web, por meio de consultas a uma máquina de busca. Esse método foi generalizado e avaliado em outros tipos de dados em [20]. Em [18], os autores apresentam as estratégias usadas para se criar um arquivo de autoridade de veículos de publicação para a área de Ciência da Computação, bem como um método de busca nesse arquivo, o qual usa um classificador associativo.

Em [12], os autores também discutem a criação de um arquivo de autoridade de veículos de publicação, o qual foi feito de forma semi-automática, usando técnicas de agrupamento (*clustering*) com posterior correção manual de dados incorretos. Arquivos de autoridade para outros atributos bibliográficos também já foram focos de pesquisa. Em [9], os autores investigaram o uso de técnicas de casamento aproximado de cadeias de caracteres para criação de arquivos de autoridade de afiliações de autores. Em [6], o foco foi nomes de editoras.

Alguns estudos avaliam a produtividade de grupos de pesquisa com base em métricas de qualidade dos veículos de publicação onde seus pesquisadores publicam [15, 21]. Desta forma, arquivos de autoridade como os desenvolvidos neste trabalho são úteis para tais estudos, pois eles ajudam a identificar o correto veículo de publicação de cada artigo.

3. PVAF

Nesta seção, são apresentadas uma breve descrição das características da versão original do PVAF, descrito em [18], e as novas funcionalidades incorporadas neste trabalho.

3.1 PVAF Original

A Figura 2 ilustra o diagrama do PVAF. Os módulos em cinza constituem a sua versão original, criada por [18]. O arquivo de autoridade foi criado por meio da coleta de dados da Web, a partir de fontes como DBLP, ACM Digital Library, IEEE Computer Science Digital Library, Qualis Capes e outras, e continha somente veículos de publicação da área de Ciência da Computação. Os dados coletados passaram por um processo de limpeza e normalização, e então foi aplicado um algoritmo de agrupamento, baseado no método K-Nearest Neighbors (Knn) [23], para identificar as diferentes referências para o mesmo veículo de publicação. O resultado gerou um arquivo de autoridade contendo mais de 11,5 mil referências para mais de 5,5 mil veículos de publicação distintos, entre periódicos, conferências e workshops.

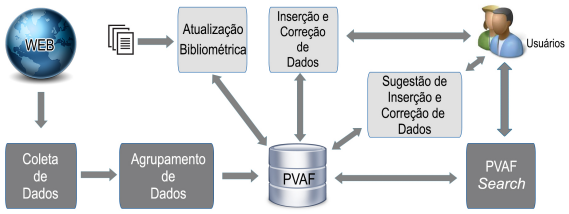

Figura 2: PVAF original (módulos em cinza) e suas novas funcionalidades (módulos em tonalidade clara)

Em [18], também foi desenvolvido um método de consulta ao PVAF, denominado PVAF Search na Figura 2. Ele é baseado em uma estratégia de aprendizagem supervisionada que usa os dados do PVAF para treinar um classificador, cujo modelo gerado forma um conjunto de regras de associação [1] para identificar veículos de publicação. As regras de associação são da forma $\mathcal{X} \rightarrow pv_i$, onde $\mathcal{X}$ é um conjunto de tokens (palavras) e pv_i é um veículo de publicação (ex: $\{SBSI\} \rightarrow pv_1$). Na fase de predição, ele gera conjuntos de tokens a partir da cadeia de caracteres a ser pesquisada, casa-os com os antecedentes das regras do modelo gerado na fase de treino e, usando um esquema de votação, decide sobre a predição.

3.2 Novas Funcionalidades

Em sua versão original, o PVAF cobria apenas a área de Ciência da Computação, e não possuía ferramentas capazes de mantê-lo atualizado. Neste trabalho, novas funcionalidades foram adicionadas ao PVAF de forma a torná-lo um sistema gerenciador de informação sobre veículos de publicação científica. Esse sistema foi denominado PVAF Manager, e suas novas funcionalidades e conjuntos de dados são descritos a seguir. Os novos módulos estão ilustrados na Figura 2 em tonalidade clara.

3.2.1 Expansão dos Dados

Com o objetivo de expandir o número de veículos de publicação contidos no banco de dados do PVAF, foram importados os dados sobre periódicos do Qualis Capes de todas as

áreas do conhecimento cobertos pela plataforma Sucupira[6]. De acordo com esses dados, cada periódico possui o ISSN (International Standard Serial Number), o título e o estrato Qualis para cada área de avaliação. Existem 49 áreas, e cada uma define seus critérios para atribuição de valores para o índice Qualis de seus periódicos de interesse. A avaliação e definição desses índices é feita anualmente. Assim, o PVAF precisa armazenar, para cada veículo de publicação, o valor do Qualis para cada área e para cada ano. A Figura 1 exemplifica a classificação Qualis do periódico "Logic Journal of the IGPL" para cada área que o avaliou.

Após a expansão dos dados, o PVAF passou a armazenar 29.452 veículos de publicação distintos, sendo 25.279 periódicos, 2.783 conferências, 1.346 workshops e 44 magazines. O número total de variações de nomes é 56.679, incluindo siglas e títulos correntes, títulos abreviados e títulos antigos.

3.2.2 Atualização Bibliométrica

Além do Qualis Capes, o PVAF também armazena o índice bibliométrico Fator de Impacto para periódicos [10]. O Fator de Impacto é calculado anualmente desde 1975 para os periódicos listados no JCR (Journal Citation Reports)[7]. O Fator de Impacto de um periódico em um determinado ano é calculado como o número de citações recebidas naquele ano por artigos publicados no periódico durante os dois anos anteriores, dividido pelo número total de artigos publicados naquele periódico durante os dois anos anteriores.

Foram desenvolvidos módulos de atualização para cada um dos índices bibliométricos, Qualis Capes e Fator de Impacto. Para o Qualis, foram criados módulos separados para atualização de periódicos e para atualização de conferências. Os dados são obtidos da planilha extraída da plataforma Sucupira para cada ano de avaliação. Na atualização, os veículos de publicação ainda não existentes no PVAF são automaticamente inseridos, bem como as novas variações de títulos e siglas de veículos de publicação existentes. Para o Fator de Impacto, o índice anual de cada periódico existente no PVAF é obtido do JCR.

3.2.3 Inserção e Correção de Dados

O PVAF foi criado por um processo automático a partir da integração de dados de diversas fontes. Na versão original, esses dados também passaram por uma conferência manual de especialistas. No entanto, devido ao grande volume de dados, alguns dados incorretos podem ter passados desapercebidos. Dados incorretos também podem ter sido introduzidos durante a expansão para outras área do conhecimento, bem como durante a atualização bibliométrica. Assim, tornou-se necessária a elaboração de um módulo para correção de dados, o qual permite a exclusão de veículos de publicação ou a alteração de valores de seus atributos. Além disso, esse módulo também permite a inserção de novos veículos de publicação, o que habilita a expansão de seus dados.

A inserção e correção de dados só é permitida para usuários administradores do sistema, como uma forma de restringir a introdução de novos dados incorretos no PVAF.

3.2.4 Sugestão de Inserção e Correção de Dados

A comunidade científica tem maior capacidade de perceber dados incorretos do PVAF do que seus administradores. Assim, foi criado um módulo para que usuários possam sugerir inserções e correções de dados. Durante uma consulta usando o módulo PVAF Search, o usuário pode (i) sugerir correções nos dados do veículo de publicação, (ii) sugerir inserção de um novo veículo de publicação, caso ele não o encontre nos resultados de suas consultas ou (iii) simplesmente alertar que o resultado não está de acordo com o esperado.

As sugestões de usuários ficam armazenadas no banco de dados do PVAF até que um usuário administrador as revise e decida se aceita ou não cada uma das sugestões. Caso aceite, os dados são atualizados no PVAF e passam a ser visíveis para as consultas seguintes.

3.2.5 Gerenciamento de Acesso

O acesso às funcionalidades de atualização bibliométrica, inserção e correção de dados é feita somente por meio de autenticação no sistema, usando um usuário especial, conhecido como administrador. O acesso para consultas ao PVAF, bem como à funcionalidade de sugestão, é livre, não exigindo o cadastro de um usuário.

4. DETALHES DE IMPLEMENTAÇÃO

4.1 Banco de Dados

Em sua versão original, os dados do PVAF foram armazenados em um arquivo no formato XML, devido a sua natureza de dados semi-estruturados [8]. O único atributo com valor obrigatório em um veículo de publicação é o seu título, que pode ser escrito de forma abreviada. Os demais atributos podem ter zero ou mais valores. Para facilitar a atualização de dados, nesta versão, esses dados passaram a ser armazenados em um banco de dados relacional. Foi escolhido o Sistema Gerenciador de Bancos de Dados MySQL [7], por ser um software popular, gratuito e eficiente para esse tipo de aplicação.

Com a expansão para outras áreas do conhecimento, o esquema de dados também foi expandido. A Figura 3 mostra o diagrama relacional do novo PVAF. A tabela central, *publication_venue*, armazena um registro para cada veículo de publicação distinto. As demais tabelas estão relacionadas a essa tabela central, e armazenam zero ou mais valores para cada veículo de publicação. Um valor obrigatório para o título é forçado na aplicação.

É importante notar que esse diagrama constitui o modelo conhecido como floco de neve da modelagem multidimensional [14]. Ele facilita as consultas OLAP (*Online Analytical Processing*) em um data warehouse, o que abre novas oportunidades para o seu uso futuro.

4.2 Painel Administrativo

Com o objetivo de implementar as funcionalidades propostas neste trabalho, foi desenvolvido um painel administrativo, agrupando as funções. A Figura 4 mostra a arquitetura desse painel. Os módulos estão organizados em camadas, e as setas indicam a comunicação entre eles. Além do banco de dados contendo o arquivo de autoridade (Banco de Dados PVAF), foi criado um outro banco de dados (Banco de Dados PVAF-Dashboard) para armazenar dados administrativos. A seguir, uma breve descrição de cada módulo:

- Interface Web: fornece o meio de interação entre os usuários e o PVAF;

[6]https://sucupira.capes.gov.br/
[7]http://wokinfo.com/products_tools/analytical/jcr

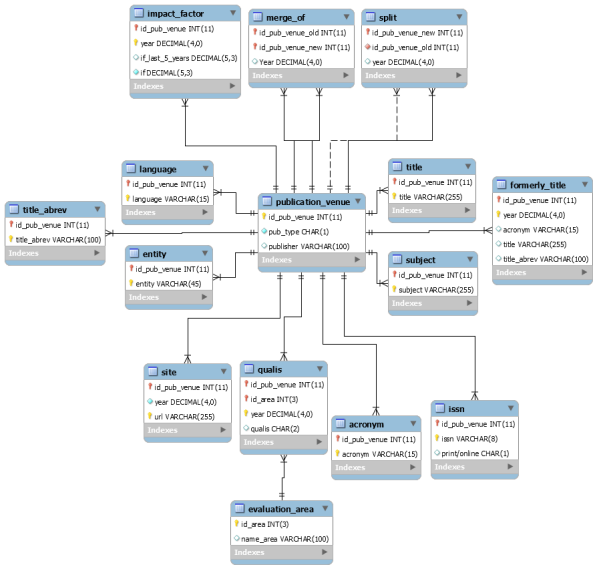

Figura 3: Diagrama Relacional do PVAF

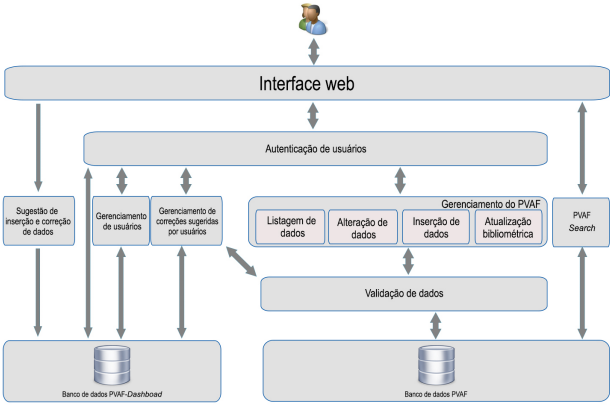

Figura 4: Arquitetura do Painel Administrativo

- Autenticação de usuários: verifica a identidade do usuário a partir do e-mail e da senha fornecidos. O módulo acessa o banco de dados PVAF-Dashboard, verifica a autenticidade dos dados e cria uma sessão de acesso ao PVAF, caso o usuário seja autenticado. Essa autenticação é exigida para se acessar as funções administrativas do PVAF;

- Gerenciamento de usuários: permite a criação e exclusão de usuários, e a manutenção de seus dados;

- Gerenciamento de correções sugeridas por usuários: permite que um usuário administrador verifique as inserções e correções sugeridas por usuários e aceite ou não cada uma delas. As sugestões são inicialmente armazenadas no banco de dados PVAF-Dashboard, e em caso de aceite, a alteração é validada e armazenada no banco de dados do PVAF. Caso a sugestão seja recusada, ela é excluída do banco de dados;

- Gerenciamento do PVAF: concentra as operações de listagem e alterações diretas de dados no PVAF. Cada requisição é submetida ao módulo de validação antes de ser efetuada no banco de dados do PVAF;

- Validação de dados: verifica a consistência das alterações e inserções solicitadas. Por exemplo, inserção duplicada de dados e fornecimento de valores obrigatórios. Caso ocorra alguma inconsistência, um erro é retornado ao módulo anterior em forma de exceção;

- PVAF Search: efetua consultas ao PVAF. O método de consulta foi proposto e implementado por [18]. Neste trabalho, o módulo foi adaptado para se conectar ao banco de dados MySQL, em vez do arquivo XML, e retornar os novos dados adicionados ao PVAF. Na inicialização, o módulo lê o banco de dados e carrega os dados em uma estrutura de índice invertido [3], a qual é usada para gerar o modelo de treinamento do classificador associativo responsável por resolver as consultas;

- Sugestão de inserção e correção de dados: recebe as sugestões dos usuários e as armazena no banco de dados PVAF-Dashboard. Elas são posteriormente analisadas por meio do módulo de gerenciamento descrito acima.

4.3 Atualização Bibliométrica

Foram desenvolvidos algoritmos para a atualização dos índices bibliométricos Qualis Capes e Fator de Impacto. A seguir, são descritos os detalhes desses algoritmos.

4.3.1 Atualização do Qualis Capes

O Algoritmo 1 mostra a implementação da atualização do índice Qualis Capes para periódicos. O algoritmo recebe como entrada o ano da classificação e o arquivo contendo o estrato Qualis de cada periódico em cada área de avaliação. A notação *p.atributo* indica o valor do atributo lido do arquivo de classificação para a instância corrente do periódico (ex: *p.issn* se refere ao valor do atributo *issn* do periódico a ser atualizado) e $PVAF.tabela$ indica o conjunto de dados contido na tabela no banco de dados do PVAF, de acordo com o diagrama da Figura 3 (ex: $PVAF.title$ se refere ao conjunto de dados da tabela *title*).

Algorithm 1: Atualização do índice Qualis Capes para Periódicos

Require: Arquivo com a classificação (arq)
Require: Ano da classificação (ano)
Ensure: PVAF atualizado
1: Leia (arq)
2: **for each** periódico $p \in arq$ **do**
3: **if** $p.issn \in PVAF.issn$ **then**
4: **if** $p.titulo \notin \{PVAF.title \cup PVAF.title\_abbrev \cup PVAF.formerly\_title\}$ **then**
5: Insere($p.issn, p.titulo, PVAF.title$)
6: **end if**
7: **else**
8: Insere($p, J, PVAF.publication\_venue, PVAF.title, PVAF.issn$)
9: **end if**
10: **if** $p.area \notin PVAF.evaluation\_area$ **then**
11: Insere($p.area, PVAF.evaluation\_area$)
12: **end if**
13: Insere($p.issn, p.qualis, p.area, ano, PVAF.qualis$)
14: **end for**

O algoritmo verifica se já existe no PVAF uma entrada com o ISSN do periódico a ser atualizado (Linha 3). Caso não exista, uma nova entrada é adicionada ao PVAF com os dados do periódico (Linha 8). Se o periódico já existe, uma nova variação de título é inserida, se ela ainda não existe (Linhas 4 e 5). Se a área de avaliação ainda não existe no PVAF, uma nova entrada é também inserida (Linhas 10 a 12). Finalmente, a classificação Qualis do periódico para o ano em questão é inserido ou atualizado (Linha 13).

Para a atualização do Qualis de conferências, a tarefa é mais complexa, pois não existe um identificador único para conferências e workshops. Na indexação do Qualis, a sigla de cada conferência é única, mas isso não ocorre no PVAF, onde um veículo de publicação pode ter zero, uma ou mais siglas, e uma sigla pode ser usada por mais de um veículo de publicação. Sendo assim, o algoritmo precisa verificar a similaridade entre siglas e títulos para identificar o veículo de publicação a ser atualizado.

Algorithm 2: Atualização do índice Qualis Capes para Conferências

Require: Arquivo com a classificação (arq)
Require: Ano da classificação (ano)
Ensure: PVAF atualizado
1: Leia (arq)
2: **for each** conferência $c \in arq$ **do**
3: $id \leftarrow 0$
4: $S \leftarrow id\_pub\_venue$ tal que $c.sigla \in PVAF.acronym$
5: **if** $S \neq \emptyset$ **then**
6: // sigla existe no PVAF
7: **for each** $s \in S$ **do**
8: $T \leftarrow \{PVAF.title \cup PVAF.title\_abbrev \cup PVAF.formerly\_title\}$ tal que $id\_pub\_venue = s$
9: **for each** $t_i \in T$ **do**
10: $sim_{t_i} \leftarrow$ SimilaridadeJaccard(t_i,c.título)
11: **end for**
12: **end for**
13: $sim \leftarrow sim_{t_i}$ tal que $sim_{t_i} > sim_{t_j} \forall j \neq i$
14: **if** $sim \geq 0.8$ **then**
15: $id \leftarrow id\_pub\_venue$ de t_i
16: **if** $c.titulo \neq t_i$ **then**
17: Insere($id, c.titulo, PVAF.title$)
18: **end if**
19: **end if**
20: **else**
21: **if** $c.titulo \in \{PVAF.title \cup PVAF.title\_abbrev \cup PVAF.formerly\_title\}$ **then**
22: $id \leftarrow id\_pub\_venue$ associado ao título encontrado
23: Insere($id, c.sigla, PVAF.acronym$)
24: **end if**
25: **end if**
26: **if** $id == 0$ **then**
27: // conferência não encontrada no PVAF
28: $id \leftarrow$ Insere($c, C, PVAF.publication\_venue, PVAF.title, PVAF.acronym$)
29: **end if**
30: Insere($id, p.qualis$, 'Ciência da Computação',$ano, PVAF.qualis$)
31: **end for**

O algoritmo pesquisa no PVAF a relação das conferências que possuem a mesma sigla da conferência c a ser atualizada (Linha 4). Se existe alguma, ele compara seus títulos com o título de c, usando o Coeficiente de Similaridade de Jaccard [13]. A conferência cujo título tenha o maior valor de similaridade é escolhida como sendo a conferência relacionada a c no PVAF, desde que essa similaridade seja maior ou igual a 0.8. O novo título é inserido no PVAF, se ele ainda não existe (Linhas 7 a 19). O valor 0.8 foi obtida por meio de avaliação experimental.

O Coeficiente de Similaridade de Jaccard é calculado da seguinte forma. Seja $|s_i|$ o número de tokens (palavras) na cadeia de caracteres s_i, e $s_i \cap s_j$ a cadeia formada pela interseção dos tokens em ambas as cadeias de caracteres. Analogamente, $s_i \cup s_j$ é a cadeia formada pela união dos tokens nas duas cadeias de caracteres. O Coeficiente de Similaridade de Jaccard $J(s_i, s_j)$ entre as cadeias de caracteres é dado pela Equação 1.

$$J(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \quad (1)$$

onde $|s_i \cap s_j|$ e $|s_i \cup s_j|$ são os números de tokens em cada cadeia de caracteres composta.

Ainda no Algoritmo 2, se a sigla da conferência a ser atualizada não é encontrada no PVAF, é verificado se o título exato existe (Linha 21). Se sim, a conferência associada ao título é escolhida como sendo a conferência relacionada a c no PVAF, e a sua sigla é inserida (Linhas 22 a 24).

Assim, uma conferência é considerada existente no PVAF se sua sigla existe e seu título possui alta similaridade com algum título da conferência no PVAF, ou se a sigla não existe, mas seu título casa exatamente com um título no PVAF. Se nenhuma dessas condições é satisfeita, a conferência é considerada nova e inserida no PVAF (Linhas 26 a 29). A conferência nova ou existente tem sua classificação Qualis atualizada (Linha 30). O PVAF armazena o Qualis de conferências somente para a área de Ciência da Computação.

4.3.2 Atualização do Fator de Impacto

Para se atualizar o índice Fator de Impacto, é necessário coletar da Web os dados do JCR de cada ano. Para isso, foi desenvolvido um *wrapper* [16], que é um programa especializado que identifica dados de interesse e os mapeia para um formato adequado. Neste caso, foram coletados o ISSN, o valor do Fator de Impacto do ano corrente e dos últimos 5 anos.

Para se fazer a atualização do PVAF, foi implementado o Algoritmo 3. Para cada periódico do PVAF, o algoritmo verifica, usando seu ISSN, se ele está indexado no JCR. Se sim, o valor do Fator de Impacto do ano corrente e dos últimos 5 anos é inserido no PVAF, na tabela *impact_factor*.

5. AVALIAÇÃO EXPERIMENTAL

Com o objetivo de validar os algoritmos de importação de dados e de consulta e verificar a cobertura dos dados do PVAF, foram feitas algumas avaliações experimentais, como apresentado a seguir.

5.1 Validação da Importação

Para validar os algoritmos de importação de dados, foram utilizadas as seguintes bases de dados: Qualis de conferência de Ciência da Computação de 2012, Qualis de periódicos de

Algorithm 3: Atualização do índice Fator de Impacto para Periódicos

Require: Arquivo do JCR (arq)
Require: Ano da classificação (ano)
Ensure: PVAF atualizado
1: Leia (arq)
2: **for each** periódico $p \in PVAF$ **do**
3: **if** $p.issn \in arq.issn$ **then**
4: Insere($p.issn, arq.fatorImpacto, ano$, $arq.fatorImpacto5anos$, $PVAF.impact\_factor$)
5: **end if**
6: **end for**

todas as áreas de 2014 e 2015, e o Fator de Impacto de acordo com o JCR de 2015. Para avaliar se os dados foram importados corretamente, após as importações, foi feita uma verificação manual no banco de dados do PVAF usando uma amostra de 20% dos veículos de publicação de cada base de dados. A Tabela 1 mostra alguns dados sobre essas bases de dados.

Base de Dados	Registros	Veic. pub. distintos	Amostra
Qualis Conf-2012	1.703	1.703	341
Qualis Per-2014	44.585	13.977	2.796
Qualis Per-2015	65.126	20.469	4.094
JCR-2015	13.862	8.776	1.756

Tabela 1: Número de registros, número de veículos de publicação distintos e tamanho da amostra de avaliação para cada base de dados.

Foi verificado na avaliação manual que todos os registros da amostra avaliada foram importados corretamente, de acordo com o algoritmo proposto. Entretanto, durante a importação, algumas inconsistências foram encontradas. A principal delas é a existência de números de ISSN fora do padrão de 8 dígitos nos arquivos do Qualis. O programa de importação gera um arquivo de registro (*log*) de inconsistências, o qual pode ser verificado posteriormente para correções manuais.

5.2 Avaliação da Cobertura do PVAF

A maioria dos dados do PVAF vem do Qualis Capes, e este indexa os periódicos e conferências nos quais a comunidade brasileira publica seus artigos. Entretanto, existem muitos outros veículos de publicação que não estão presentes no PVAF. Uma ampla avaliação de cobertura do PVAF demanda um esforço manual muito grande. Neste trabalho, foi feita uma avaliação restrita de cobertura, de acordo com a seguinte metodologia. Foram escolhidas seis áreas distintas do conhecimento para serem avaliadas. Para cada uma delas, foram escolhidas aleatoriamente as listas de publicações de 15 professores das Universidades de Stanford e Harvard, disponíveis nos sites dessas duas renomadas universidades americanas, exceto para a área de Ciências Agrárias, onde foi usada a Universidade de Delaware em vez de Harvard, por possuir mais cursos nessa área. Para cada publicação, foi extraída a cadeia de caracteres correspondente ao veículo de publicação e usada como consulta ao PVAF. Foram, então, avaliados o número de veículos de publicação encontrados no PVAF.

A Tabela 2 mostra o resultado. Para cada área do conhecimento, são apresentados o número de veículos de publicação distintos encontrados nos currículos, o percentual deles encontrados no PVAF (cobertura), o número de variações de nomes dos veículos de publicação e o percentual de erros de pesquisa, respectivamente. O percentual geral de cobertura do PVAF foi de $86,18\%$, sendo Medicina a área com maior cobertura e História, a menor.

5.3 Validação da Consulta

Para validar o algoritmo de consulta do PVAF, foi avaliado o número de erros cometidos ao retornar o veículo de publicação desejado. Cada cadeia de caracteres distinta extraída dos currículos dos pesquisados na avaliação de cobertura, conforme apresentado na seção anterior, foi usada como consulta ao PVAF e verificado, manualmente, se o resultado retornado era realmente o veículo de publicação desejado. O resultado está mostrado na coluna "Erros de Pesquisa" da Tabela 2. Foram considerados somente os veículos de publicação existentes no PVAF. A coluna "Variações de nomes" dessa mesma tabela mostra o número de cadeias de caracteres usadas como consulta.

O resultado mostra que o algoritmo de consulta do PVAF possui baixas taxas de erros, com uma média de $1,97\%$. A área com maior percentual de erros foi a Biologia e a menor, História, com nenhum erro.

6. CONCLUSÕES

Este trabalho apresentou os detalhes do desenvolvimento do PVAF Manager, um sistema gerenciador de informações sobre veículos de publicação científica. O sistema gerencia um arquivo de autoridade com variações de nomes de periódicos e conferências. O arquivo de autoridade também mantém outras informações sobre veículos de publicação, incluindo índices bibliométricos como o Qualis Capes e o Fator de Impacto.

Neste trabalho, o banco de dados de veículos de publicação do PVAF foi expandido para todas as áreas do conhecimento indexadas no Qualis da Capes, totalizando mais de 29 mil veículos de publicação distintos, com mais de 56 mil variações de nomes. Foram desenvolvidos algoritmos de importação de dados e atualização dos índices bibliométricos, os quais necessitam tratar a ambiguidade de nomes, bem como opções para a correção manual de dados por usuários gerenciadores. Além disso, está disponível um opção para que usuários possam sugerir correções e novos dados para o PVAF.

Foi feita uma avaliação experimental para validar os algoritmos de importação de dados e de consulta e verificar a cobertura dos dados do PVAF. Como resultado, foram identificadas algumas inconsistências nos dados do Qualis, e coletados automaticamente dados para que as mesmas possam ser corrigidas. O método de consulta demonstrou ser eficaz, com uma taxa de erro de apenas $1,97\%$. A cobertura geral do PVAF foi de $86,18\%$.

Como trabalhos futuros, está sendo desenvolvido um método de auto-treinamento do algoritmo de consulta ao PVAF, o qual permitirá que novos veículos de publicação e novas variações de nomes sejam incorporadas automaticamente ao modelo de treinamento de aprendizagem de máquina, sem que o sistema precise parar para ser retreinado. Além disso,

Área do Conhecimento	# Veic. publicação	Cobertura (%)	# Variações de nomes	Erros de Pesquisa (%)
Administração	147	70,75	154	0,65
Biologia	245	91,43	297	3,37
Ciência da Computação	295	84,41	407	1,97
Ciências Agrárias	215	91,63	248	1,61
História	57	64,91	58	0,00
Medicina	264	92,05	306	1,96
Geral	**1.223**	**86,18**	**1.470**	**1,97**

Tabela 2: Resultados da cobertura e da validação de consultas no PVAF para seis áreas do conhecimento.

será desenvolvida uma ferramenta que permite a comparação de grupos de pesquisa, gerando estatísticas sobre a qualidade de suas publicações.

Acknowledgements

Trabalho parcialmente suportado pelo projeto FAPEMIG CEX-APQ-01834-14 e bolsas individuais do CNPq e da UFLA.

7. REFERÊNCIAS

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago, Chile, 1994.

[2] L. Auld. Authority control: An eight-year review. *Library Resources & Technical Services*, 26:319–330, 1982.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval: The Concepts and Technology behind Search*. Addison-Wesley Professional, 2011.

[4] R. Bennett, C. Hengel-Dittrich, E. T. O'Neill, and B. B. Tillett. VIAF (virtual international authority file): Linking die deutsche bibliothek and library of congress name authority files. In *Proceedings of the World Library and Information Congress: 72nd IFLA General Conference and Council*, Seoul, Korea, August 2006.

[5] T. Berners-Lee. Linked Data, 2006. https://www.w3.org/DesignIssues/LinkedData.html. Accessed March, 2017.

[6] L. S. Connaway and T. J. Dickey. Publisher names in bibliographic data. *Library Resources and Technical Services Journal*, 55(4):182–194, 2011.

[7] P. DuBois. *MySQL*. Addison-Wesley Professional, 5th edition, 2013.

[8] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Pearson, 7th edition, 2016.

[9] J. C. French, A. L. Powell, and E. Schulman. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 51(8):774–786, 2000.

[10] E. Garfield. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93, 2006.

[11] M. Hallo, S. Luján-Mora, A. Maté, and J. Trujillo. Current state of linked data in digital libraries. *Journal of Information Science*, 42(2):117–127, 2016.

[12] N. Houssos, C. Paschou, I.-O. Stathopoulou, K. Stamatis, and D. Hardouveli. Implementing citation management and report generation value-added services over oai-pmh compliant repositories. In *Proceedings of the 5th International Conference on Open Repositories*, Madrid, Spain, July 2010.

[13] P. Jaccard. étude comparative de la distribuition florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[14] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley, 3rd edition, 2013.

[15] A. H. F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 4(2):135–145, 2008.

[16] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, 2002.

[17] D. Lee, J. Kang, P. Mitra, C. L. Giles, and B.-W. On. Are your citations clean? *Communications of the ACM*, 50(12):33–38, December 2007.

[18] D. A. Pereira, E. E. B. da Silva, and A. A. A. Esmin. Disambiguating publication venue titles using association rules. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries*, pages 77–86, London, UK, Setember 2014.

[19] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, and A. H. F. Laender. Using web information for creating publication venue authority files. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 295–304, Pittsburgh, USA, June 2008. ACM New York, NY, USA.

[20] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. F. Laender, and M. A. Gonçalves. A generic web-based entity resolution framework. *Journal of the American Society for Information Science and Technology*, 62(5):919–932, May 2011.

[21] S. Ribas, B. Ribeiro-Neto, E. de Souza e Silva, A. H. Ueda, and N. Ziviani. Using reference groups to assess academic productivity in computer science. In *Proceedings of the 24th International Conference on World Wide Web*, pages 603–608, New York, NY, USA, 2015. ACM.

[22] VIAF: The virtual international authority file, 2017. http://viaf.org/. Accessed in March, 2017.

[23] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, third edition, 2011.