HEIDER ALVARENGA DE JESUS

ENRIQUECENDO UM ARQUIVO DE AUTORIDADE DE VEÍCULOS DE PUBLICAÇÃO COM INFORMAÇÕES EXTRAÍDAS DA WEB

LAVRAS - MG

2015

HEIDER ALVARENGA DE JESUS

ENRIQUECENDO UM ARQUIVO DE AUTORIDADE DE VEÍCULOS DE PUBLICAÇÃO COM INFORMAÇÕES EXTRAÍDAS DA WEB

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Banco de Dados e Engenharia de Software, para a obtenção do título de Mestre.

Orientador

Dr. Denilson Alves Pereira

LAVRAS - MG

2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Jesus, Heider Alvarenga de.
Enriquecendo um arquivo de autoridade de veículos de publicação com informações extraídas da Web / Heider Alvarenga de Jesus. – Lavras : UFLA, 2015.
78 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2015.
Orientador(a): Denilson Alves Pereira.
Bibliografia.

1. Arquivo de Autoridade. 2. Veículo de Publicação. 3. Extração de Informação. 4. Página Web. 5. Máquina de Busca. I. Universidade Federal de Lavras. II. Título.

HEIDER ALVARENGA DE JESUS

ENRIQUECENDO UM ARQUIVO DE AUTORIDADE DE VEÍCULOS DE PUBLICAÇÃO COM INFORMAÇÕES EXTRAÍDAS DA WEB

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Banco de Dados e Engenharia de Software, para a obtenção do título de Mestre.

APROVADA em 13 de Julho de 2015.

Dr. Anderson Almeida Ferreira UFOP

Dr. Guilherme Tavares de Assis UFOP

Dr. Denilson Alves Pereira

Orientador

LAVRAS - MG

2015

Dedico esta dissertação aos meus pais, Nelson e Maria pelo amor e apoio durante toda a caminhada. Às minhas irmãs, Tamar e Elen, pela amizade e carinho. Aos meus parentes, amigos e colegas que sempre estiveram ao meu lado confiando e apoiando.

AGRADECIMENTOS

Agradeço, primeiramente a Deus, pela minha vida, oportunidades concedidas e por sempre ser fiel.

Agradeço aos meus pais, Nelson e Maria, e as minhas irmãs, Tamar e Elen, que sempre me apoiaram e, com muito amor, deram condições para chegar até aqui.

Agradeço à Universidade Federal de Lavras, ao Departamento de Ciência da Computação e ao Programa de Pós-Graduação em Ciência da Computação da UFLA, pela estrutura oferecida e pela oportunidade de realização do Mestrado.

Agradeço à Capes pelo apoio financeiro que possibilitou a realização do Mestrado.

Agradeço ao meu orientador Prof. Denilson pela confiança, orientação, conselhos e disponibilidade.

Agradeço a todos os professores, colegas, secretárias e funcionários do Departamento de Ciência da Computação da UFLA pelo conhecimento repassado, pelas ajudas e pelo apoio.

Agradeço a minha família, amigos e colaboradores que souberam me entender e que ofereceram o ombro amigo para a realização deste trabalho.

E a todos que contribuíram, torceram e acreditaram em minha conquista. Muito obrigado!

RESUMO

Arquivos de autoridade mantêm registros de entidades e são normalmente utilizados por bibliotecas digitais na elaboração de ferramentas de desambiguação de nomes de autores ou títulos de veículos de publicação. Um arquivo de autoridade com informações detalhadas e consistentes sobre veículos de publicação permite a melhoria de tais ferramentas. Neste trabalho, objetivou-se enriquecer um arquivo de autoridade de veículos de publicação da área de Ciência da Computação. A proposta é obter informações adicionais para complementar esse arquivo de autoridade já existente, extraindo informações automaticamente de páginas da Web, obtidas por meio de consultas a uma máquina de busca. A abordagem contempla os passos para submissão de consultas, classificação dos documentos obtidos por elas e extração de informações dos documentos relevantes. A classificação das páginas é uma tarefa importante neste trabalho. Duas abordagens foram implementadas e avaliadas experimentalmente: classificação baseada apenas em conteúdo e classificação baseada em gênero e conteúdo. A primeira obteve melhores resultados para páginas de conferências. Das páginas relevantes, foram extraídos dados como ano, número da edição e data, além do nome e sigla, em busca de alguma variante desconhecida na forma de escrita. Os experimentos realizados demonstram bons resultados na coleta de informações de conferências, permitindo-se traçar um histórico de realização das mesmas, com dados como ano de suas edições e mudanças de nomes.

Palavras-chave: Arquivo de Autoridade. Veículo de Publicação. Extração de Informação. Classificação de Documentos. Máquina de Busca.

ABSTRACT

Authority files maintain entity registries and are generally used by digital libraries for elaborating disambiguation tools for author names or titles of publishing venues. An authority file with detailed and consistent information on publication venues allows the improvement of such tools. This work has the objective of enriching an authority file of Computer Science publication venue. The proposal is of obtaining additional information in order to complement this already existing authority archive, by automatically extracting information from web pages, obtained by means of consultations to a research engine. The approach contemplates the steps for submitting consultations, classifying documents and extracting information of relevant documents. The classification of the pages is an important task in this work. Two approaches were implemented and experimentally evaluated: classification based only on content, and classification based on gender and content. The first obtained the best results for page conference. From the relevant pages, we extracted data such as year, edition number and date, in addition to name and abbreviation, seeking an unknown variant in written form. The experiments conducted demonstrate good results in the collection of conference information, allowing us to trace the record of performing the same, with data such as edition year and name change.

Keywords: Authority File. Publication Venue. Information Extraction. Document Classification. Search Engine.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACM	*Association for Computing Machinery*
API	*Application Programming Interface*
CBC	Classificação Baseada em Conteúdo
CBG	Classificação Baseada em Gênero
CBGC	Classificação Baseada em Gênero e Conteúdo
DBLP	*Digital Bibliography & Library Project*
EC1	Estratégia de Classificação 1
EC2	Estratégia de Classificação 2
IE	*Information Extraction*
IEEE	*Institute of Eletrical and Eletronics Engineers*
IHF	*Inverse Host Frequency*
IR	*Information Retrieval*
JIF	*Journal Impact Factor*
OCLC	*Online Computer Library Center*
SIGMETRICS	*ACM Conference on Measurement and Modeling of Computer Systems*
SJR	*SCImago Journal & Country Rank*
SVM	*Support Vector Machines*
VIAF	*Virtual International Authority File*
VLDB	*Very Large Data Bases Conference*

SUMÁRIO

1 INTRODUÇÃO

Arquivos de autoridade mantêm registros de entidades e podem ser utilizados por ferramentas para diversos fins como, por exemplo, para desambiguação de nomes de autores ou veículos de publicação e para medir a qualidade de um grupo de pesquisadores. Com base nos trabalhos de Pereira et al. (2008) e Pereira, Silva e Esmin (2014), pode-se dizer que manter arquivos de autoridade atualizados e com informações consistentes não é uma tarefa trivial, inclusive muito extensa para ser realizada manualmente. Neste trabalho propõe-se uma abordagem para o enriquecimento de um arquivo de autoridade de conferências da área de Ciência da Computação com informações extraídas automaticamente de páginas da Web, obtidas por meio de consultas à uma máquina de busca. Neste capítulo, são apresentadas a contextualização e motivação, a proposta, a justificativa, os objetivos gerais e específicos e a estrutura do trabalho.

1.1 Contextualização e motivação

A qualidade de uma publicação científica pode ser aferida, usando-se índices bibliométricos, tais como o Journal Impact Factor (JIF)[1], o SCImago Journal & Country Rank (SJR)[2] e o Qualis Capes[3]. Entretanto, para se utilizar essas medidas com precisão, é importante identificar corretamente o título do veículo de publicação de citações extraídas de artigos científicos.

Sistemas de indexação bibliográfica de bibliotecas digitais trabalham com uma vasta quantidade de dados e constantemente lidam com registros inconsistentes. Citações a veículos de publicações como periódicos, confe-

[1] <http://wokinfo.com/products_tools/analytical/jcr/>
[2] <http://www.scimagojr.com/>
[3] <http://qualis.capes.gov.br/webqualis/principal.seam>

rências e *workshops* muitas vezes possuem erros ortográficos, variações na escrita e abreviações, que acabam tornando a busca e a recuperação mais difíceis (PEREIRA et al., 2008).

Uma maneira encontrada pelas bibliotecas digitais para tentar contornar o problema de identificar formas variantes se referindo a uma mesma entidade é a utilização de Arquivos de Autoridade. De acordo com Auld (1982), um arquivo de autoridade contém variações de escrita utilizadas para uma determinada entidade, que pode ser um campo bibliográfico, por exemplo. Porém, a criação de arquivos de autoridade não é uma tarefa fácil.

No trabalho de Pereira, Silva e Esmin (2014), foi construído um arquivo de autoridade de veículos de publicação da área de Ciência da Computação. Esse arquivo é formado por um conjunto de registros com informações sobre títulos e siglas correntes, títulos e siglas antigas, dentre outras informações sobre veículos de publicação. Na Figura 1 mostra-se um exemplo de um registro no arquivo de autoridade. Uma ferramenta para consulta ao arquivo de autoridade está disponível em http://pvaf.dcc.ufla.br.

```
<pub-venue>
   <id>22</id>
   <entity>IEEE</entity>
   <acronym>CCC</acronym>
   <title>Conference on Computational Complexity</title>
   <title>IEEE Conference on Computational Complexity</title>
   <title>Annual IEEE Conference on Computational Complexity</title>
   <title-formerly>Structure in Complexity Theory Conference</title-formerly>
   <acronym-formerly>CoCo</acronym-formerly>
   <pub-type>C</pub-type>
   <qualis-estrato>B1</qualis-estrato>
</pub-venue>
```

Figura 1 Exemplo de um registro no arquivo de autoridade criado por Pereira, Silva e Esmin (2014)

O registro da Figura 1 está no formato XML. A *tag* id armazena o identificador único do registro, a *tag entity*, a entidade responsável pelo

veículo de publicação, *acronym*, a sigla, *title*, os títulos, *title-formerly*, o título antigo, *acronym-formerly*, a sigla antiga, *pub-type*, o tipo do veículo de publicação (C para conferência, J para periódico ou W para workshop) e o *qualis-estrato*, a classificação do veículo no índice bibliométrico da Capes.

Além de variações de escrita de nomes, arquivos de autoridade também podem armazenar outros dados referentes às entidades, permitindo assim uma melhor utilização em recuperação de informação. Informações sobre os números das edições de uma conferência, anos de ocorrência, anos das mudanças de nomes (caso ocorreram) e URLs, por exemplo, não existem no arquivo de autoridade proposto por Pereira, Silva e Esmin (2014) e seriam importantes para o seu enriquecimento.

1.2 Proposta deste trabalho

A hipótese levantada neste trabalho é que as informações para enriquecer o arquivo de autoridade podem ser obtidas da Web, por meio de consultas a uma máquina de busca. Os maiores desafios são identificar os documentos relevantes nos resultados das consultas e extrair os dados contidos neles, desde que não exista um padrão para escrita de páginas da Web.

Este trabalho apresenta uma abordagem para realizar a coleta de páginas da Web relacionadas a um veículo de publicação, identificar sua provável página oficial e extrair informações adicionais sobre ele, de forma a enriquecer o arquivo de autoridade criado por Pereira, Silva e Esmin (2014). Em razão das peculiaridades das páginas dos diferentes tipos de veículos de publicação, como periódicos, conferências e *workshops*, diferentes estratégias precisam ser elaboradas, adaptando-as à realidade de cada tipo de veículo

de publicação. Neste trabalho, foram abordadas estratégias para a coleta de informações sobre conferências.

A abordagem é baseada nos seguintes passos. Primeiro, submetem-se consultas à uma máquina de busca e, por meio dos resultados dessas consultas, obtém-se as URLs e coletam-se suas páginas HTML. Em seguida, as páginas passam por uma classificação, com o objetivo de selecionar aquelas relevantes para a extração de dados. São consideradas relevantes as páginas oficiais do veículo de publicação em cada uma de suas edições. Uma estratégia usada na classificação é a abordagem de classificação baseada em gênero e conteúdo (ASSIS; LAENDER; GONÇALVES, 2008), onde o classificador analisa os termos referentes à estrutura da página (gênero) e, posteriormente, os termos relativos ao veículo de publicação em si (conteúdo). Em outra estratégia, utilizou-se apenas a classificação baseada em conteúdo. E finalmente, um extrator baseado em expressões regulares varre as páginas classificadas como relevantes em busca do número da edição, título, sigla e data de ocorrências. Então, a partir dos dados extraídos de cada conferência, traça-se um histórico de suas edições, incluindo a identificação das mudanças de nomes, quando existem.

As principais contribuições deste trabalho são: (a) a proposta de uma abordagem que coleta, identifica e extrai informações de páginas oficiais de conferências, (b) um conjunto de experimentos que demonstram a eficácia da abordagem proposta e (c) a coleta de um conjunto de dados que enriqueceram o arquivo de autoridade de Pereira, Silva e Esmin (2014).

1.3 Justificativa

É possível comparar grupos de pesquisa por meio da qualidade dos veículos de publicação em que seus membros publicam. Como dito anteriormente, existem índices bibliométricos que avaliam a qualidade dos veículos de publicação. Um conjunto de citações é extraído do currículo de pesquisadores e, para um uso efetivo dos índices bibliométricos, é necessário identificar corretamente o veículo de publicação. Como o título de um veículo de publicação pode vir escrito de maneiras diferentes, ferramentas de desambiguação de títulos de veículos de publicação podem utilizar arquivos de autoridade para realizar tal tarefa (PEREIRA; SILVA; ESMIN, 2014).

A ideia é que, com um arquivo de autoridade enriquecido com informações mais detalhadas sobre veículos de publicação, possa haver uma melhoria dessas ferramentas, que podem utilizar as informações para garantir uma melhor consistência nos resultados. Por exemplo, na ferramenta de desambiguação de veículos de publicação, além de verificar as variações de escrita do nome do veículo, também podem ser verificados o volume, número, edição ou a data, dependendo do tipo de veículo de publicação. Além disso, outros tipos de ferramentas poderão utilizar os dados do arquivo de autoridade. Por exemplo, um aplicativo que gere uma linha do tempo dos veículos de publicação, ou uma plataforma de consulta com as informações dos veículos.

Obter informações automaticamente na Web e de forma consistente não é uma tarefa trivial. Tais informações podem estar disponíveis na própria página oficial do veículo de publicação, em bibliotecas digitais, em páginas de divulgação de eventos científicos, em redes sociais ou outras páginas que replicam as informações divulgadas. Uma maneira de garantir a

consistência das informações é extraí-las diretamente da página oficial do veículo de publicação.

Coletar informações de páginas não oficiais ou de bibliotecas digitais pode gerar problemas, no sentido de que as informações replicadas da página oficial podem conter erros ou estarem desatualizadas. A coleta de informações sobre veículos de publicação diretamente nas bibliotecas digitais também pode apresentar alguns problemas como:

- embora pareça que haja um padrão na disposição das informações, muitas vezes elas não estão estruturadas e não estão disponíveis diretamente para coleta;
- seria necessário desenvolver coletores específicos para cada biblioteca digital e eles seriam sensíveis a qualquer mudança no formato das páginas;
- as bibliotecas digitais contêm informações incompletas e, muitas vezes, incorretas;
- mesmo que as grandes bibliotecas digitais fossem incluídas no processo de coleta de informações, nem todos os veículos de publicação seriam encontrados.

Já na Web aberta, as chances de se encontrar dados sobre qualquer veículo de publicação podem ser maiores. O desafio é construir um coletor eficiente e eficaz para tal tarefa.

1.4 Objetivos geral e específicos

O objetivo geral deste trabalho é enriquecer um Arquivo de Autoridade de Veículos de Publicação da área de Ciência da Computação com

informações extraídas automaticamente da Web, por meio de consultas a uma máquina de buscas.

Para que tal objetivo seja alcançado, é necessário atender aos seguintes objetivos específicos:

- desenvolver um algoritmo para submeter consultas à uma máquina de buscas e coletar as páginas de seus resultados;
- desenvolver um classificador para identificar as páginas relevantes;
- desenvolver um algoritmo para extrair dados das páginas relevantes;
- agrupar os algoritmos desenvolvidos em um arcabouço capaz de extrair informações do conjunto de dados coletados;
- avaliar experimentalmente a proposta;
- enriquecer um arquivo de autoridade com os dados coletados.

1.5 Organização do trabalho

Este documento está organizado da seguinte forma. No Capítulo 2, é apresentado o Referencial Teórico com uma breve discussão sobre os assuntos abordados, bem como os trabalhos relacionados e suas contribuições para este trabalho. A Abordagem Proposta é apresentada no Capítulo 3, seguida da Avaliação Experimental no Capítulo 4, que descreve e avalia os experimentos, apresentando uma discussão sobre os resultados obtidos. A Conclusão e Trabalhos Futuros são apresentados no Capítulo 5.

2 REFERENCIAL TEÓRICO

Para o desenvolvimento deste trabalho, é necessário apresentar alguns conceitos e definições concernentes a algumas áreas do conhecimento. O trabalho abrange conceitos que envolvem bibliotecas digitais, desambiguação de entidades, arquivos de autoridade, consultas em máquinas de buscas, classificação de documentos e extração de informação, que são apresentados com mais detalhes nas subseções a seguir. Neste capítulo, também são apresentados os trabalhos relacionados e suas contribuições para este trabalho.

2.1 Bibliotecas digitais

Existem várias formas e tipos de bibliotecas. A evolução das tecnologias de informação e comunicação, principalmente com a expansão da Internet, permitiu o surgimento de mudanças na forma de gerar, distribuir, colecionar, acessar e utilizar a informação e as bibliotecas. Hoje, as tecnologias digitais e suas aplicações fazem parte da vida das pessoas e permitiram o surgimento das bibliotecas digitais (MA; CLEGG; O'BRIEN, 2006).

Bibliotecas digitais são repositórios de dados que possuem uma coleção de documentos e informações armazenados em formatos digitais acessíveis por meios eletrônicos. A acessibilidade pela Internet permite a quebra de barreiras de distância, um fator limitante das bibliotecas físicas. De acordo com Ma, Clegg e O'Brien (2006), biblioteca digital é uma nova forma de gerenciar o registro do conhecimento e do patrimônio cultural.

As bibliotecas digitais estão se tornando cada vez mais utilizadas em todo o mundo e não há dúvida de que o futuro da criação de conhecimento e do compartilhamento de informações encontram-se nas redes eletrônicas

(KAVULYA, 2007). O desenvolvimento das bibliotecas digitais provocou uma profunda revolução na distribuição de publicações científicas e acadêmicas (ROSA; SHMORGUN; LAMAS, 2012).

Na área de Ciência da Computação, existem bibliotecas digitais bem consolidadas e utilizadas pelos pesquisadores. ACM – *Association for Computing Machinery*[4], CiteSeer^{X}[5], DBLP – *Digital Bibliography & Library Project*[6], IEEE Xplore – *Institute of Electrical and Electronics Engineers*[7] e BDBComp – Biblioteca Digital Brasileira de Computação[8] são algumas delas.

Na Figura 2, exibe-se a página principal da biblioteca digital IEEE Xplore, onde pode ser visualizada a ferramenta de busca do site, que permite a realização de buscas em toda a base de publicações científicas armazenadas na biblioteca digital. As ferramentas de busca das bibliotecas digitais geralmente possuem diversos filtros e configurações que podem ser aplicados às buscas com o intuito de refiná-las para obter resultados mais satisfatórios. Além da ferramenta de busca, o site da IEEE Xplore, por exemplo, permite a navegação pelas conferências, livros, periódicos, dentre outros. Muitas bibliotecas digitais também possuem ferramentas que auxiliam o usuário a encontrar o que está sendo procurado.

Citações são muito importantes em bibliotecas digitais. Uma citação é uma coleção de informações bibliográficas pertencentes a um determinado documento científico. De acordo com Lee et al. (2007), tanto os usuários de bibliotecas digitais como os pesquisadores usam citações, seja para bus-

[4] <http://dl.acm.org/>
[5] <http://citeseerx.ist.psu.edu/>
[6] <http://dblp.uni-trier.de/>
[7] <http://ieeexplore.ieee.org/>
[8] http://www.lbd.dcc.ufmg.br/bdbcomp/

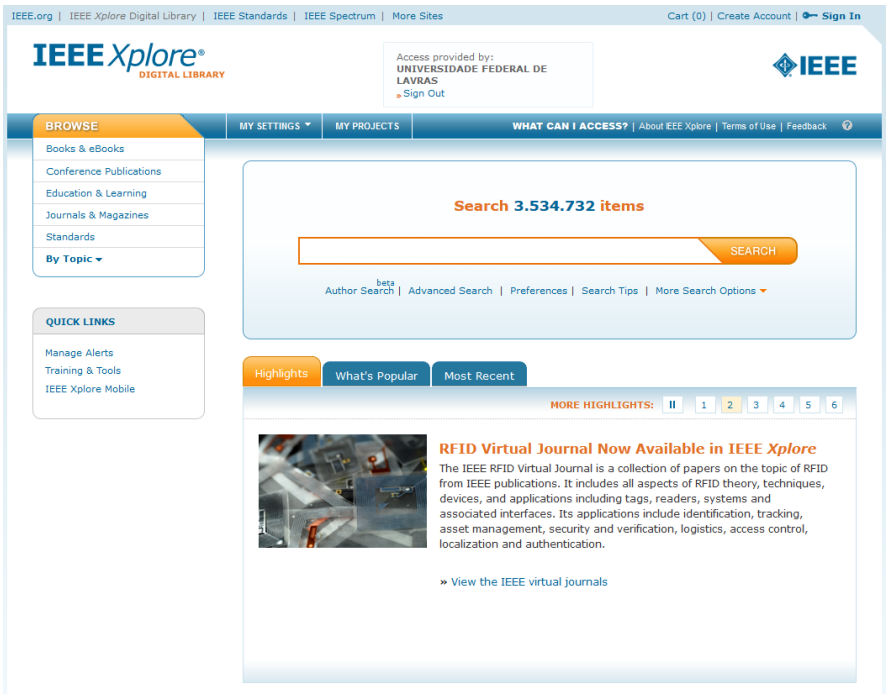

Figura 2 Página principal da biblioteca digital IEEE Xplore

car informações nas bibliotecas digitais ou para avaliar o impacto de um determinado artigo. Os autores afirmam que podem haver muitas citações inconsistentes e/ou desatualizadas armazenadas nas bibliotecas digitais.

Existem vários desafios para manter a consistência desses dados: erros na entrada de dados, formato das citações, falta de padrões, software de coleta imperfeito, nomes de autores ambíguos e abreviações de nomes dos veículos de publicação são exemplos. Manter a consistência desses dados não é uma tarefa trivial (LEE et al., 2007).

Bibliotecas digitais tradicionais mantêm arquivos de autoridade e trabalham para manter sua consistência e, com isso, reduzir problemas como a ambiguidade de nomes de autores ou as variações de nomes de veículos de publicação (PEREIRA et al., 2008).

2.2 Desambiguação de entidades

Em bibliotecas digitais, usualmente ocorre o problema de ambiguidade de entidades. A ambiguidade ocorre quando existem múltiplas entidades com o mesmo nome (polissemia) ou diferentes variações de nomes para a mesma entidade (sinônimos).

Lee (2007) apresentam dois problemas relacionados à ambiguidade de nomes de autores em bibliotecas digitais: a citação misturada (*mixed citation*) e a citação dividida (*split citation*). O problema da citação misturada ocorre quando citações de diferentes autores são misturadas, o que ocorre em decorrência de múltiplos autores terem o mesmo nome ou a grafia de seus nomes serem muito similares. Por exemplo, o autor "H. Jesus" pode se referir a "Heider Jesus" ou "Hugo Jesus", duas pessoas diferentes. O problema da citação dividida ocorre quando citações do mesmo autor são divididas, como se fossem de autores diferentes. Por exemplo, a ocorrência do nome do autor "Heider Alvarenga de Jesus" pode aparecer em múltiplas publicações com diferentes abreviações de nomes, como "Heider Jesus", "H. Jesus", ou "H. A. Jesus", ou com erro de grafia como "Heider Lavarenga".

Um problema similar também pode ocorrer com nomes de veículos de publicação. Na Tabela 1, apresenta-se um exemplo particular da variedade de referências distintas à conferência SIGMETRICS (*ACM Conference on Measurement and Modeling of Computer Systems*). Essas referências correspondem não somente a formas distintas de escrita do nome desta conferência, mas também a nome de instâncias que se referem a suas edições com nomes antigos.

Ferramentas de desambiguação de entidades visam a reduzir diversos problemas como os citados acima. Vários métodos de desambiguação foram

Tabela 1 Exemplos de referências distintas à conferência SIGMETRICS

String do Veículo de Publicação
"ACM Conference on Measurement and Modeling of Computer Systems"
"ACM SIGMETRICS Conference"
"ACM Conference on Measurement and Modelling of Computing Systems"
"ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems"
"ACM joint international conference on Measurement and modeling of computer systems"
"ACM international conference on Measurement and modeling of computer systems"
"international joint conference on Measurement and modeling of computer systems"
"ACM conference on Simulation, measurement and modeling of computer systems"
"ACM conference on Computer performance modeling measurement and evaluation"
"ACM conference on Measurement and evaluation"

propostos e podem ser encontrados na literatura (FERREIRA; GONÇALVES; LAENDER, 2012; PEREIRA et al., 2008, 2009; PEREIRA; SILVA; ESMIN, 2014).

2.3 Arquivos de autoridade

De acordo com Auld (1982), um arquivo de autoridade mantém a correspondência entre variantes de escritas usadas para se referir a uma mesma entidade. Um arquivo de autoridade é um índice de registros de autoridade, onde cada registro, representando uma entidade, é composto por um nome a ser usado como rótulo da entidade (nome canônico) e uma lista de rótulos variantes também usados para se referir à entidade (PEREIRA; SILVA; ESMIN, 2014).

Uma das principais iniciativas para criação de arquivos de autoridade é o projeto VIAF[9] (*Virtual International Authority File*), que combina vários arquivos de autoridade em um único serviço hospedado pela OCLC[10] (*Online Computer Library Center*). Objetiva-se, neste serviço, reduzir o custo e aumentar a utilidade das bibliotecas de arquivos de autoridade, combinando e vinculando arquivos de autoridade amplamente utilizados, tornando a informação disponível na Web (ONLINE COMPUTER LIBRARY CENTER - OCLC, 2015). A proposta deste trabalho não visa a contribuir diretamente para o projeto VIAF, mas trabalha com um arquivo de autoridade específico que possui registros de veículos de publicação de Ciência da Computação. Segundo Pereira, Silva e Esmin (2014), o projeto VIAF não possui registros detalhados específicos da área de Ciência da Computação, especialmente de conferências.

Connaway e Dickey (2011) apresentam um projeto onde foi construído um arquivo de autoridade contendo as maiores editoras (*publishers*) do mundo. Os autores usaram o prefixo do ISBN para agrupar registros bibliográficos por editora, resultando em um banco de dados com milhares de variações de nomes de editoras e dados sobre elas.

French, Powell e Schulman (2000) propuseram um método para gerar um arquivo de autoridade para um conjunto de afiliações de autores. O objetivo do método é identificar e agrupar (*clustering*) referências distintas a uma mesma instituição. O processo de agrupamento requer uma função que compara a distância entre diferentes *strings*. Os autores relataram o efeito de diferentes funções de distância. Quando o processo de agrupamento é concluído, para cada grupo é associado uma *string* canônica, dada

[9] <https://viaf.org/>
[10] <http://www.oclc.org/>

pela *string* mais frequente no grupo. As técnicas descritas pelos autores são eficazes e podem ser usadas na construção automática de arquivos de autoridade. Apesar disso, em algumas situações, ainda requer uma intervenção manual para resolver alguns casos.

Pereira et al. (2008) e Pereira, Silva e Esmin (2014) propuseram estratégias para se construir arquivos de autoridade de veículos de publicação; maiores detalhes sobre esses trabalhos serão dados posteriormente nas Seções 2.4 e 2.8, respectivamente.

2.4 O arquivo de autoridade de Pereira, Silva e Esmin (2014)

Uma das contribuições do trabalho de Pereira, Silva e Esmin (2014) foi a construção de um arquivo de autoridade de veículos de publicação da área de Ciência da Computação. Este arquivo é formado por um conjunto de registros em que cada um representa um veículo de publicação. Cada registro é composto por um cabeçalho para ser usado como rótulo do veículo de publicação e uma lista de variações de nomes que podem ser usadas para referenciar o veículo de publicação, chamadas referências cruzadas.

O arquivo de autoridade contém informações sobre periódicos, conferências e *workshops*. Para cada veículo de publicação, ele armazena as formas variantes dos títulos e siglas atuais, títulos e siglas antigos, fusões e algumas informações extras, tais como ISSN, editora, idioma, assunto e índices bibliométricos, como o Qualis Capes e o Fator de Impacto. Na Figura 3, apresenta-se um exemplo das informações contidas em um registro do arquivo de autoridade. Neste exemplo, o registro contém a sigla, 3 variações do título, um título antigo, 4 variações de fusão de títulos, 4 de fusão de siglas, o tipo e o estrato qualis do veículo de publicação. O primeiro

elemento *pv.title* é considerado o rótulo do veículo de publicação, enquanto as outras variações do título, bem como o título antigo, fusões de títulos, siglas, siglas antigas e fusões de siglas são as chamadas referências cruzadas.

pv.sigla	AAMAS
pv.título	International Conference on Autonomous Agents and Multiagent Systems
	International Joint Conference on Autonomous Agents and Multiagents Systems
	International Joint Conference on Autonomous Agents and Multiagent Systems
pv.título-antigo	international conference on Advanced Agent Technology
pv.fusão-de-título	International Conference on Autonomous Agents
	Workshop on Agent Theories, Architectures, and Languages
	International Conference on Multiagent Systems
	International Workshop on Agent Theories, Architectures, and Languages
pv.fusão-de-sigla	AGENTS
	ATAL
	ICMAS
	AA
pv.tipo	C
pv.estrato-qualis	A2

Figura 3 Exemplo de um registro do arquivo de autoridade de veículos de publicação

O arquivo de autoridade contém 11.592 referências a 5.524 veículos de publicação (1.937 periódicos, 2.227 conferências e 1.344 *workshops*). Uma interface de busca está disponível em http://pvaf.dcc.ufla.br.

O arquivo de autoridade foi criado por meio da coleta de dados de bibliotecas digitais e instituições que organizam *rankings* de qualidade de veículos de publicação na área de Ciência da Computação.

Os autores criaram coletores que identificam dados de interesse, fazem a extração e armazenam os dados em um banco de dados. Uma grande dificuldade encontrada pelos autores foi decorrente das fontes de dados não serem estruturadas, além de conter informações incompletas e, algumas vezes, incorretas.

Até então, o arquivo de autoridade pode conter, para cada registro, várias siglas e títulos, mas não possui uma correspondência entre qual sigla está associada a qual título, tanto para os dados correntes como para

os antigos. Ele também não possui informações de quando ocorreram as mudanças de nomes. Para obter tais informações, é necessário ter fontes de dados mais completas, que podem ser obtidas na Web.

Neste trabalho, objetivou-se obter dados adicionais da Web para enriquecer o arquivo de autoridade de veículos de publicação já criado. São utilizados as siglas e os títulos atuais e as siglas e os títulos antigos para a geração de *strings* de busca para realização das consultas.

2.5 Consultas a máquinas de buscas

Uma máquina de busca Web ou motor de busca (*search engine*) é um *software* que permite aos usuários pesquisar e recuperar documentos da Web por meio de consultas, de acordo com suas necessidades de informação (TUMER; SHAH; BITIRIM, 2009). Ela é um sistema de recuperação de informação capaz de oferecer uma representação, armazenamento, organização e acesso aos itens de informação como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia, dentre outros. A representação e a organização desses itens de informação provêm aos usuários um fácil acesso à informação de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2011).

Uma visão geral sobre a arquitetura do *software* de um sistema de recuperação de informação (Figura 4) auxilia o entendimento dos procedimentos executados por uma máquina de busca Web.

De acordo com Baeza-Yates e Ribeiro-Neto (2011), a processo de coleta consiste em montar uma coleção de documentos privados ou coletados da Web. Um coletor (*crawler*) é responsável por coletar os documentos. A coleção de documentos é armazenada em um repositório central e, no pro-

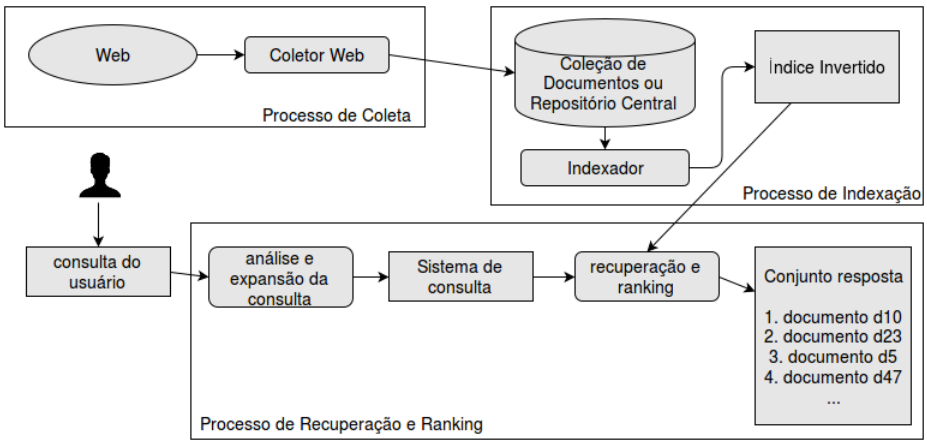

Figura 4 Arquitetura de um sistema de recuperação de informação, baseado em Baeza-Yates e Ribeiro-Neto (2011)

cesso de indexação, os documentos desse repositório são indexados gerando um índice invertido. O processo de recuperação e *ranking* utiliza esse índice invertido para gerar o *ranking*, que é utilizado para uma rápida recuperação e posicionamento. Com isso, pode-se realizar a recuperação de documentos que satisfazem uma consulta do usuário.

As máquinas de busca normalmente disponibilizam uma API (*Application Programming Interface*) para que outros programas possam submeter consultas de forma automática a elas. Como elas possuem um grande repositório de dados e sofisticados algoritmos de recuperação de informação, muitos trabalhos utilizam dessas APIs para recuperar informações de interesse.

A máquina de busca do Google, por exemplo, possui a *Google Search Engine API*[11], que permite a utilização do *Google Search* em suas aplicações. Embora seja possível utilizar essa API gratuitamente, ela possui uma limitação quanto ao número de consultas realizadas e a forma como são

[11] <https://developers.google.com/web-search/>

realizadas. Se a API detecta um alto número de buscas de uma máquina e/ou que um algoritmo esteja executando as consultas automaticamente em um curto período de tempo, ela passa a bloquear as consultas. É possível contornar tal limitação regulando-se o algoritmo que efetua as consultas a fazê-las entre períodos aleatórios de tempo, o que acaba estendendo um pouco o tempo gasto na etapa de submissão de consultas.

Vários trabalhos utilizaram consultas a máquinas de buscas para diversos propósitos. O trabalho de Silva et al. (2009) utiliza os registros de metadados de um documento existente em uma biblioteca digital, cujo texto completo não está disponível, para realizar consultas em máquinas de busca à procura da URL correspondente ao texto completo. Para isso, os autores apresentaram um estudo sobre o processo de utilização de diferentes estratégias de consulta para artigos de conferências de Ciência da Computação aplicados a diferentes máquinas de busca e considerando diferentes necessidades e perfis do usuário. Com a realização dos experimentos, os autores concluíram que o Scholar é a melhor alternativa para essa tarefa. A segunda alternativa é o Google, sendo alcançado pelo Yahoo! com um desempenho equivalente, nos cenários em que os usuários consideraram apenas documentos que podem ser acessados gratuitamente. Para essas três máquinas de buscas, as consultas sem utilizar aspas foram mais eficientes que as consultas que as utilizaram. As aspas em uma consulta limita a máquina a pesquisar os termos exatamente na ordem em que estão na consulta. CiteSeer e MSN não se mostraram boas alternativas para essa tarefa em particular.

Pereira et al. (2009) utilizam informações da Web para desambiguar nomes de autores. A proposta do trabalho é coletar informações de citações de entrada e submeter consultas a uma máquina de busca Web, objetivando

encontrar o currículo e páginas Web contendo publicações de autores ambíguos. Este método aproveita o aumento do uso da Web por pesquisadores científicos para publicar seus trabalhos, bem como os grandes repositórios implementados pelos motores de busca da Web. Os resultados obtidos pelos autores indicam um grande ganho na qualidade na desambiguação, quando comparada com métodos não-supervisionados, e está empatado estatisticamente com métodos de aprendizagem supervisionados, os quais requerem rotulamento feito por humanos e possuem um tempo de treinamento caro. O método proposto pelos autores também é interessante para ser usado em conjunto com outras estratégias de desambiguação em um procedimento de agrupamento hierárquico.

O trabalho de Pereira et al. (2008) também envolve consultas à máquinas de busca. Mais detalhes são apresentados na Seção 2.8.

2.6 Classificação de documentos

É comum as pessoas buscarem padrões em dados. Sejam padrões de comportamento animal, de crescimento da cultura nas lavouras, no clima e até padrões na opinião de eleitores.

Profissionais de diversas áreas do conhecimento trabalham na ideia de que padrões podem ser encontrados em dados que permitam, automaticamente, identificar, validar e ser usado para fazer previsões. A quantidade de dados cresce, a cada dia, e fica praticamente impossível para um ser humano analisar todos esses padrões. Mineração de dados trata de resolver os problemas encontrados para analisar os dados presentes em bancos de dados, onde os dados são armazenados eletronicamente e a busca é automatizada pelo computador (WITTEN; FRANK; HALL, 2011).

A classificação é umas das técnicas utilizadas em Mineração de Dados. O processo de classificação pode ser supervisionado ou não-supervisionado.

Na classificação supervisionada, o processo é dividido em duas etapas: treinamento e teste. No treinamento, um modelo (classificador) é gerado a partir da análise de um conjunto de dados de treinamento, cujos dados estão previamente rotulados. Na etapa de teste, o modelo construído é aplicado ao conjunto de teste, formado por classes desconhecidas. O modelo gerado deve ser capaz de descrever e distinguir automaticamente as classes de dados cujos rótulos são desconhecidos (KAMBER; PEI, 2012).

Já na classificação não-supervisionada, conhecida como agrupamento ou *clustering*, o rótulo de cada instância de treinamento não é conhecido e o número ou conjunto de classes a serem aprendidas pode não ser conhecido com antecedência (KAMBER; PEI, 2012).

Jalil, Kamarudin e Masrek (2010) descrevem algumas das técnicas utilizadas para classificação de dados. Dentre elas, o *Suport Vector Machines* (SVM), que foi utilizada em uma das estratégias na etapa de classificação desenvolvida neste trabalho.

O SVM, introduzido por Cortes e Vapnik (1995), constitui-se de um método de espaço vetorial para problemas de classificação binária. Os termos indexados compõem um espaço t-dimensional, onde os dados são representados por pontos. Dadas as representações vetoriais para os documentos, a ideia é buscar um hiperplano que pode ser usado para separar os elementos em duas classes. O hiperplano, que é gerado a partir de um conjunto de dados de treino, divide o espaço em duas regiões, de forma que todos os documentos de uma classe fiquem em uma região e todos os docu-

mentos da outra classe fiquem em outra região. Uma vez que o modelo foi gerado, um novo documento pode ser classificado computando-se sua posição em relação ao hiperplano (BAEZA-YATES; RIBEIRO-NETO, 2011). Para ilustrar, considere o exemplo de um espaço bi-dimensional cujos pontos dos dados de treinamento são linearmente separáveis (podem ser separados por uma linha simples), como mostra a Figura 5. Os pontos representados por losangos e triângulos representam os documentos de duas classes distintas. Dentre todas as linhas que separam os documentos em duas classes, a linha L_2 maximiza as distâncias do documento mais próximos de cada classe e constitui o melhor hiperplano de separação. Note que a linha L_1 fornece uma alternativa pior neste caso porque suas distâncias dos documentos mais próximos nas classes c_a e c_b são menores; já a linha L_3 sequer separa as duas classes.

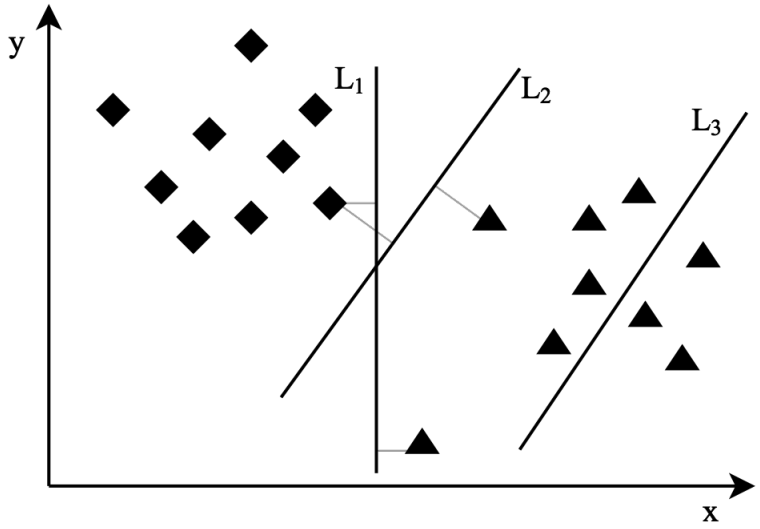

Figura 5 SVM: L_1 separa as classes, mas com um intervalo pequeno, L_2 separa com intervalo máximo entre as classes e L_3 não separa. Adaptado de Weinberg (2012)

2.7 Extração de informação

A cada dia, há mais dados em forma de texto na forma eletrônica do que nunca, mas muitos desses dados são ignorados. Nenhum humano pode ler, entender e sintetizar tamanha quantidade de texto. Informações perdidas – e oportunidades perdidas – motivaram, nos últimos anos, a exploração de estratégias de gerenciamento da informação. As estratégias mais comuns são a Recuperação de Informação (*Information Retrieval – IR*), Filtragem de Informação (*Information Filtering*) e também a Extração de Informação (*Information Extraction – IE*) (COWIE; LEHNERT, 1996).

De forma geral, enquanto sistemas de IR coletam material útil de uma vasta quantidade de matéria-prima, os sistemas de IE transformam a matéria-prima, refinando e reduzindo-as no núcleo da informação do texto original. Os sistemas de IR coletam os documentos relevantes, enquanto os sistemas de IE recebem essa coleção de documentos e os transformam em informações mais organizadas e mais fáceis de serem analisadas. Os sistemas de IE isolam fragmentos de texto relevantes e extraem as informações relevantes desses fragmentos (COWIE; LEHNERT, 1996).

O núcleo de um sistema de IE é um extrator, que processa texto, ignorando palavras e frases irrelevantes, e tenta encontrar entidades e relações entre elas. Por exemplo, um extrator pode mapear a sentença "Minas Gerais é o estado mais promissor do Brasil" para a tupla relacional (*Minas Gerais, Estado do, Brasil*), que pode ser representada por alguma linguagem formal (ETZIONI et al., 2008).

Em textos muito longos, é necessário ter um conhecimento muito amplo para extrair precisamente essas tuplas. Existem técnicas que permitem obtê-las, como a codificação baseada no conhecimento direto (um

especialista humano insere expressões regulares ou regras), a aprendizagem supervisionada (um especialista humano fornece exemplos de treinamento já rotulados) e a aprendizagem não-supervisionada (o sistema rotula automaticamente seus próprios exemplos) (ETZIONI et al., 2008).

Na Web, a maioria das informações é apresentada em páginas HTML. Elas geralmente são semiestruturadas e a informação requerida para um contexto pode estar dispersa em diferentes documentos. É difícil analisar um grande volume de documentos semiestruturados apresentados em páginas Web e tomar decisões baseadas nessas análises. O trabalho apresentado por Abburu e Babu (2013) propõe um *framework* para um sistema que extrai a informação de várias fontes e prepara um relatório baseado no conhecimento obtido por meio de análises. Ele integra um coletor Web, um extrator de informação e tecnologias de mineração de dados para uma melhor análise das informações e para ajudar nas decisões a serem tomadas. O *framework* é aplicável em diversos domínios como fábricas, vendas, turismo e outros. A coleta é feita por meio de um *crawler* que vai coletando e extraindo as URLs contidas nas páginas.

Abburu e Babu (2013) utilizaram para a extração de dados, a ferramenta *Easy Web Extractor*[12], que extrai conteúdos de páginas Web como textos, URLs, imagens e arquivos e transformam os resultados em múltiplos formatos de saída, não requerendo programação. Os autores utilizaram o software Weka[13] para a mineração de dados. O sistema proposto permite ao usuário buscar facilmente todas as páginas Web de um determinado domínio URL, extrair os dados de todas as páginas e fazer análises nos dados extraídos usando mineração de dados.

[12]<http://www.webextract.net>
[13]<http://www.cs.waikato.ac.nz/ml/weka/>

2.8 Trabalhos relacionados

Este trabalho constitui-se de três etapas principais, sendo elas a submissão de consultas à uma máquina de busca, a classificação de documentos coletados e a extração de informação. A seguir, é apresentada uma série de trabalhos que abrangem esses temas e estão relacionados a este trabalho.

Os trabalhos desenvolvidos por Pereira et al. (2008, 2009) adotaram a estratégia de extrair informações da Web, por meio de submissão de consultas a uma máquina de busca, para criação de um arquivo de autoridade de veículos de publicação e para desambiguação de nomes de autores, respectivamente. De acordo com os autores, a estratégia de submissão de consultas a máquina de buscas obteve bons resultados. Ademais, a ideia de obter as páginas dos veículos de publicação, por meio dos resultados de consultas a uma máquina de busca, mostrou-se diretamente aplicável a este trabalho.

Pereira et al. (2008) propõem a utilização de informações disponíveis na Web para criar um arquivo de autoridade de veículos de publicação. A ideia é reconhecer e extrair referências à veículos de publicação dos trechos de texto (*snippets*) retornados pela máquina de busca. As referências a um mesmo veículo de publicação são reconciliados em um arquivo de autoridade, onde cada entidade é composta por um nome canônico para o veículo, uma sigla, o tipo de veículo (i.e., periódico, conferência ou *workshop*) e o mapeamento para várias formas de escrita do nome em citações bibliográficas. Como o objetivo deste trabalho é de enriquecer um arquivo de autoridade de veículos de publicação com informações adicionais, é necessário obter informações além das apresentadas nos *snippets*. Por isso, adotou-se a estratégia de coletar as páginas das URLs fornecidas nas respostas da

máquina de busca. Tendo coletado a página do veículo de publicação, o número de informações que podem ser extraídas aumenta consideravelmente em relação às que podem ser coletadas apenas nos *snippets*, a um custo de processamento maior.

O trabalho de Silva et al. (2009) propõe um processo para recuperar a URL de um documento para o qual existem metadados em um catálogo de uma biblioteca digital, mas um *link* para o texto completo do documento não está disponível. O processo proposto pelos autores usa os resultados de consultas submetidas a máquinas de busca Web buscando a URL correspondente ao texto completo ou algum material relacionado. Os autores estudaram a abordagem proposta em situações distintas, investigando a aplicação de diferentes estratégias de consultas em três máquinas de busca de propósito geral (Google, Yahoo! e MSN) e duas especializadas (Scholar e CiteSeer), considerando cinco cenários de usuários. Nos resultados experimentais obtidos pelos autores, eles concluíram que o processo proposto é eficaz e proporciona uma estratégia muito simples para encontrar o texto completo dos documentos catalogados em uma biblioteca digital cuja URL para o texto completo está ausente. Além disso, mostraram que dentre as máquinas de busca testadas, a Scholar é a mais eficaz para essa tarefa e, quando combinada com o Google, ganhos significativos são alcançados para todos os cenários considerados. Os autores apresentaram um estudo sobre o processo de utilização de diferentes estratégias de consulta aplicados a diversas máquinas de busca e considerando diferentes necessidades e perfis do usuário. Este estudo trouxe informações sobre o processo de consulta nas máquinas, como a geração de *strings* de busca e quais formatos obtiveram resultados satisfatórios. O formato da *string* de busca pode influenciar os

resultados da máquina de busca. Neste trabalho, a forma que as *strings* de busca foram geradas baseia-se nos resultados apresentados pelos autores.

Os trabalhos de Assis et al. (2008, 2009) e Assis, Laender e Gonçalves (2008) apresentam uma abordagem baseada em gênero para coleta de páginas da Web relacionadas a um tópico específico, podendo ser expressas em termos de gênero e conteúdo. Coletores focados coletam páginas Web relacionadas a um tópico ou interesse específico, mas alguns usuários podem não estar simplesmente interessados em algum documento sobre um tópico e sim em documentos de um determinado estilo ou gênero referente a um tópico.

Na abordagem deles, o processo de coleta pode ser expresso por dois conjuntos de termos: o primeiro descrevendo aspectos de gênero das páginas desejadas e o segundo relacionado ao assunto ou conteúdo da página. Os autores também avaliaram o impacto da seleção de um conjunto de termos gerados por um especialista humano, outro gerado por um usuário familiarizado com o tema e também uma estratégia para geração semiautomática dos termos de gênero e conteúdo. Os autores avaliaram que a geração manual de termos geralmente é suficiente para produzir bons resultados e a geração semiautomática é muito eficaz no apoio à tarefa de seleção do conjunto de termos. Por meio de um conjunto de experimentos, os autores demonstraram a eficácia, a eficiência e a escalabilidade da abordagem proposta.

Neste trabalho, utilizou-se as abordagens baseada em gênero e conteúdo e baseada apenas em conteúdo para a classificação de páginas de veículos de publicação. Como a maioria dessas páginas apresenta uma estrutura semelhante, é possível determinar um conjunto de termos de gênero, formado por palavras comuns neste tipo de página. Já os termos de con-

teúdo são formados por palavras que representam o veículo de publicação que está sendo classificado, de forma a evitar que páginas de outros veículos de publicação sejam classificadas como sendo do veículo em questão.

No trabalho de Alfred et al. (2014), os autores descrevem um *framework* robusto de recuperação de informação cuja proposta é auxiliar usuários a acessar informações relevantes eficiente e eficazmente, entregando as consultas de acordo com suas preferências. O *framework* coleta páginas HTML e executa diversos processamentos nelas, como limpeza de *tags* HTML e transformação na representação de espaço vetorial com pesos TF-IDF (*Term Frequency - Inverse Document Frequency*). Neste trabalho, a aplicação de tais tarefas é necessária para preparar os dados de entrada para o classificador baseado em gênero. Os autores também utilizam uma função *ranking* $f(q,d)$, onde q é a consulta e d o documento, para organizar os resultados baseado na consulta dos usuários. A mesma estratégia da função *ranking* foi adaptada e utilizada no classificador baseado em conteúdo neste trabalho, de forma que as páginas fiquem organizadas em ordem de relevância.

Com o crescimento exponencial das informações disponíveis na Web, requer-se urgentemente novas técnicas para discernir informações úteis das informações desnecessárias. No trabalho de Gunasundari e Karthikeyan (2012), em vez de apenas extrair o conteúdo relevante de páginas Web, os autores investigaram como remover vários padrões de ruídos delas e assim obter informações sobre o conteúdo principal. Os autores focaram na remoção de ruídos como *tags* que não estão relacionadas com informações válidas (<img>, <script> e <style>, por exemplo), barras de navegação, barras laterais, etc. Nos experimentos conduzidos neste trabalho, executou-se al-

guns métodos para remover padrões desnecessários em algumas etapas. A remoção de *tags* HTML e a extração de conteúdos específicos, como blocos de *links* e menus, foram aplicadas em alguns dos experimentos.

O arquivo de autoridade gerado no trabalho de Pereira, Silva e Esmin (2014) foi utilizado neste trabalho, conforme já descrito na Seção 2.4.

3 ABORDAGEM PROPOSTA

Neste trabalho, elaborou-se uma abordagem para enriquecer o arquivo de autoridade gerado por Pereira, Silva e Esmin (2014), que é o arquivo de autoridade base deste trabalho. A abordagem proposta submete consultas a uma máquina de busca e coleta páginas da Web, das quais são extraídas informações sobre os veículos de publicação. Este trabalho contemplou veículos de publicação do tipo conferência, que normalmente são realizadas em edições anuais, onde cada edição possui sua própria página. Na Figura 6, ilustra-se de forma simplificada a abordagem proposta.

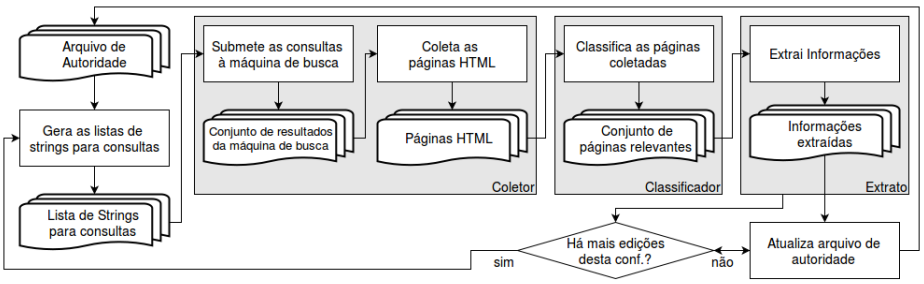

Figura 6 Fluxograma da abordagem proposta

Em razão das características das páginas Web de veículos de publicação, algumas modificações devem ser adaptadas para cada tipo de veículo (conferências, periódicos e *workshops*), para que a obtenção de dados seja satisfatória. Sendo assim, este trabalho focou no desenvolvimento de uma abordagem para a obtenção de informações sobre conferências. Uma peculiaridade das conferências, por exemplo, é o fato delas geralmente serem realizadas em edições anuais: cada edição possuindo sua própria página.

De acordo com a Figura 6, o processo inicia-se com a obtenção dos dados básicos das conferências a partir do arquivo de autoridade criado

por Pereira, Silva e Esmin (2014). Com esses dados, são geradas strings para consultas em uma máquina de busca. Inicialmente, são executadas consultas relacionadas aos dados atuais de cada conferência, usando-se o ano corrente. Depois, são buscados os anos anteriores, reduzindo-se ano a ano até que fique um determinado período sem encontrar páginas relevantes. Um classificador é utilizado para dizer se uma página é ou não relevante. Passa-se, então, a utilizar os dados sobre títulos e siglas antigos das conferências nas consultas a partir do primeiro ano em que não foi encontrada nenhuma página relevante, reduzindo-se ano a ano novamente, até que fique mais um determinado período sem encontrar nenhuma página relevante, encerrando-se a coleta para a conferência. Depois repete-se o processo para as próximas conferências.

Usando-se as páginas relevantes coletadas, é feita a extração das informações para complementar o arquivo de autoridade. E do conjunto de páginas coletadas, também é possível traçar um histórico de cada conferência com os anos em que ela ocorreu e quando houve uma troca de nome, por exemplo. Nas seções seguintes, são apresentados os detalhes de cada etapa.

3.1 Geração de *strings* para consultas

Para a etapa de geração de *strings* para consultas, recebe-se um conjunto de siglas atuais e antigas e títulos atuais e antigos das conferências, obtido por meio do arquivo de autoridade base deste trabalho. Objetivou-se, nesta etapa, construir um conjunto de combinações atuais e antigas, com as siglas e títulos das conferências unindo cada sigla com cada título, separando apenas os atuais dos antigos. A saída desta etapa serão dois conjuntos de *strings* compostos pelas combinações formadas.

Formalmente, seja $PV = \{pv_1, pv_2, ..., pv_n\}$ um conjunto de n veículos de publicação obtido do arquivo de autoridade criado por Pereira, Silva e Esmin (2014). Para cada pv_i, são obtidos os conjuntos $A = \{a_1, a_2, ..., a_o\}$, $T = \{t_1, t_2, ..., t_p\}$, $AF = \{af_1, af_2, ..., af_q\}$ e $TF = \{tf_1, tf_2, ..., tf_r\}$, onde A é um conjunto de siglas, T de títulos, AF de siglas antigas e TF de títulos antigos de um determinado veículo de publicação.

Para cada $pv_i \in PV$, são geradas duas listas de strings, $L = \{l_1, l_2, ..., l_{o \cdot p}\}$ e $LF = \{lf_1, lf_2, ..., lf_{q \cdot r}\}$. A lista L é composta por *strings* com as combinações entre os conjuntos A e T e a lista LF é composta por *strings* com as combinações entre os conjuntos AF e TF. Isso é necessário pois não existe uma correspondência entre siglas e títulos no arquivo de autoridade atual.

Considere o exemplo da Figura 7, que mostra o registro do veículo de publicação "*International Symposium on Formal Methods*" no arquivo de autoridade. Do lado esquerdo, tem-se as siglas e os títulos da conferência, e do lado direito tem-se as listas de *strings* para consultas geradas pelos dados do registro.

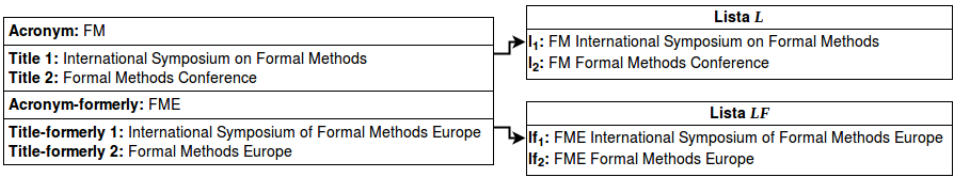

Figura 7 Exemplo de geração de *strings* para consultas

Em muitos casos, a lista L fica muito grande e com *strings* semelhantes. Em razão disso, optou-se pela utilização de uma amostra desta lista, para evitar a submissão de muitas consultas, as quais obtêm praticamente os mesmos resultados, de acordo com testes preliminares. Para isso, adotou-se

a estratégia de reduzir a lista L, deixando nela apenas a primeira combinação mais 30% das demais *strings* da lista, escolhidas aleatoriamente. Esse número foi obtido empiricamente.

Já a lista LF não é alterada, pois pode haver diferentes títulos antigos, em decorrência de mais de uma mudança na sigla ou título do veículo de publicação.

Nos casos em que o arquivo de autoridade não possui a sigla do veículo de publicação, a lista foi composta apenas com os títulos.

3.2 Submissão de consultas a uma máquina de busca

Nesta etapa, recebe-se como entrada os dois conjuntos de *strings* gerados na etapa de geração de *strings* para consultas e o valor do ano corrente (ano atual). Objetivou-se, nesta etapa, submeter consultas a uma máquina de busca, utilizando as strings dos conjuntos de entrada e obter um conjunto de respostas para cada submissão. Como saída, esta etapa fornece um conjunto de resultados contendo o título e a URL de cada página obtida pela máquina de busca.

Para cada veículo de publicação, inicia-se as consultas na máquina de busca com a lista L. Ao final de cada *string* de L, é adicionado o valor do ano corrente e as consultas são submetidas à máquina de busca. À medida que um classificador identifica a página relevante, o algoritmo decrementa o ano e submete novas consultas. Caso o classificador não encontre uma página relevante para um determinado ano, ele é marcado e a busca continua pelos anos anteriores até no máximo, 5 anos seguidos. Se nesse período nenhuma página relevante for encontrada, o algoritmo retorna ao ano marcado e passa a utilizar a lista LF. Após um período de 5 anos sem encontrar páginas

relevantes, encerram-se as consultas para a conferência e passa-se para a próxima. Esse período de 5 anos foi obtido empiricamente, após observação do comportamento das páginas pesquisadas.

Pode ocorrer de o arquivo de autoridade ter registros de conferências mais antigas e que já foram descontinuadas há mais de 5 anos. Estabeleceu-se, então, um prazo de 25 anos para se buscar a edição mais recente, que coincide com o início da Internet. Na Figura 8, ilustra-se o processo de submissão de consultas à máquina de busca.

As consultas são submetidas à máquina de busca do Google, por meio da *Google Search Engine API*[14]. A máquina de busca do Google foi escolhida por ser reconhecida atualmente como a melhor máquina de busca da Web, além de ter apresentado bons resultados no trabalho de Silva et al. (2009).

Para cada consulta submetida, são coletados os seus 10 primeiros resultados, formando um conjunto de respostas $\{g_1, g_2, ..., g_{10}\}$. Depois, os resultados das consultas, para cada ano, são unidos em um único conjunto $G = \{g_1, g_2, ..., g_n\}$.

3.3 Coleta e pré-processamento das páginas Web

Esta etapa recebe como entrada o conjunto de resultados G da máquina de busca contendo um conjunto de títulos e URLs das páginas. Objetivou-se, nesta etapa, coletar o conteúdo HTML de cada página do conjunto de resultados, armazená-la localmente e realizar um pré-processamento para preparar os dados da página para a etapa de classificação. A saída desta etapa é um conjunto de dados pré-processados prepara-

[14]https://developers.google.com/web-search/

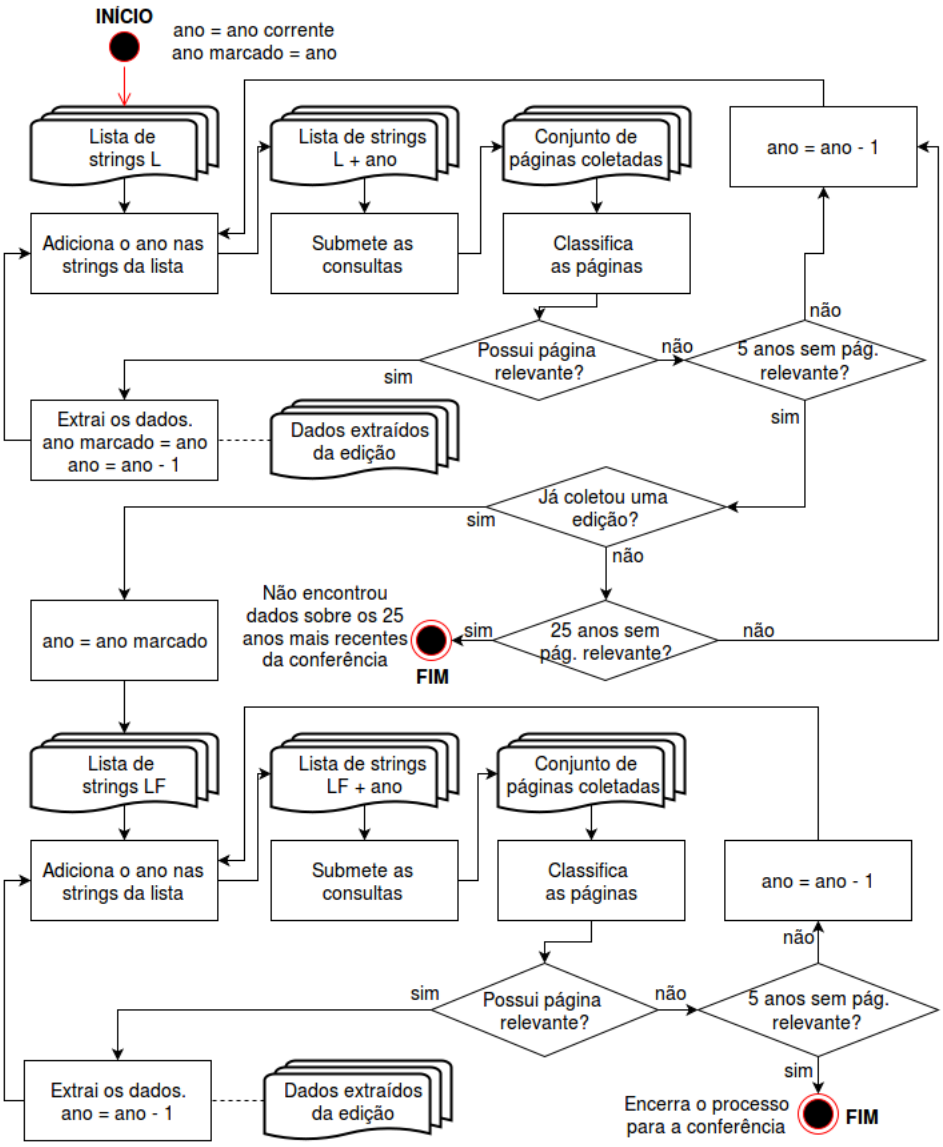

Figura 8 Fluxograma do processo de submissão de consultas para uma conferência

dos para serem utilizados na etapa de classificação, além de uma cópia local de cada página HTML coletada para ser utilizada na etapa de extração de dados.

Para cada elemento g_i do conjunto G gerado na etapa anterior, o módulo de coleta acessa a URL da página e armazena localmente o seu conteúdo HTML. URLs que remetem a arquivos que não possuam conteúdo HTML, como arquivos PDF, DOC ou PPT, por exemplo, são ignorados pelo coletor, pois não há interesse em outro tipo de conteúdo que não seja páginas HTML, uma vez que as páginas oficiais geralmente estão nesse formato.

O módulo de coleta também identifica quando há casos de páginas de redirecionamento ou compostas por *frames*. Nesses casos, ele coleta também o conteúdo das URLs que apontam para páginas adicionais e considera o conjunto de páginas coletadas como sendo referente à página original, permitindo que todo o conteúdo exibido ao usuário nesse tipo de página possa ser recuperado.

As páginas coletadas são então submetidas a um processo de limpeza, onde são removidas as *tags* HTML, as *stopwords*, os dígitos e os caracteres especiais, e os caracteres maiúsculos são convertidos em minúsculos. Com isso, é gerado um conjunto de termos, onde cada palavra da página é considerada como um *token*, que será utilizado pelo classificador baseado em gênero. Posteriormente, os arquivos HTML originais voltam a ser utilizados na extração de dados.

3.4 Classificação das páginas relevantes

Esta etapa recebe como entrada um conjunto contendo os dados coletados nas etapas anteriores. Esses dados incluem os resultados das consultas à máquina de busca, a página HTML e os dados pré-processados da página. Recebe também um conjunto de termos de gênero e outro de termos de conteúdo, a serem utilizados pelo classificador. Objetivou-se, nesta etapa,

classificar cada página coletada como sendo ou não relevante, e assim descartar as páginas que não são úteis para a abordagem. A saída desta etapa é um conjunto de páginas de conferências classificadas como relevantes.

A etapa de classificação verifica se uma página é relevante. Uma página é considerada relevante quando ela é a página oficial do veículo de publicação em questão. Para a classificação das páginas, foram utilizadas duas abordagens: uma baseada na combinação gênero e conteúdo (ASSIS et al., 2008; ASSIS; LAENDER; GONÇALVES, 2008) e outra baseada apenas no conteúdo.

Na classificação baseada em gênero, é analisada a presença dos termos mais comuns em um determinado tipo de página e, geralmente, eles definem a estrutura da página. Em páginas de conferências, por exemplo, são comuns termos de gênero como: *conference, submission, dates, call, papers, keynote, speakers, committe, accommodation, program*, dentre outros. Já os termos de conteúdo são os termos que representam o veículo de publicação específico, que está sendo classificado naquele momento. No caso das conferências, por exemplo, os termos de conteúdo são formados pela sigla, termos contidos no título e ano da conferência.

Os termos de gênero foram gerados de duas maneiras: manualmente por um especialista humano e automaticamente por um algoritmo de seleção de *features*. Os termos foram gerados manualmente por meio da observação de um conjunto de páginas de conferências, onde os termos mais comuns foram listados. Os outros termos foram gerados automaticamente no software Weka[15] por meio de um algoritmo de seleção de *features* que gera um conjunto de *features* ordenados por relevância.

[15]http://www.cs.waikato.ac.nz/ml/weka/

Na classificação baseada em gênero, são usados como atributos os termos de gênero, os quais são usados para se calcular os valores do TF-IDF (*Term Frequency - Inverse Document Frequency*) (BAEZA-YATES; RIBEIRO-NETO, 2011) nos documentos da coleção de treinamento. Os valores do TF-IDF são passados para o SVM (*Support Vector Machines*) (VAPNIK, 1995) fazer a classificação. Por ser uma técnica de classificação supervisionada, o SVM requer um treinamento prévio, com documentos classificados por um especialista. A partir desse treino, ele gera um modelo que é utilizado para classificar os demais documentos.

Na classificação baseada em conteúdo, o algoritmo analisa o título da página, sua URL e seu conteúdo. Para cada veículo de publicação, o algoritmo gera um *ranking* das páginas, conforme um conjunto de pontuações para cada página coletada. São analisadas a presença do (a) ano na URL, (b) ano no título da página, (c) sigla na URL, (d) sigla no título da página, (e) título do veículo no conteúdo da página e (f) ano no conteúdo da página, sendo considerados, apenas, os 2 mil primeiros caracteres da página como conteúdo, pelo fato de que as informações como título, sigla e ano da conferência normalmente virem no início da página. O algoritmo também garante que haja pelo menos uma menção ao título, por meio de casamento exato de termos, ou sigla do veículo de publicação e uma ao ano que está sendo buscado.

Foram adotadas duas estratégias para a classificação baseada em conteúdo. Na Estratégia de Classificação 1 (EC1), o algoritmo seleciona apenas uma página como relevante para cada conferência. Já na Estratégia de Classificação 2 (EC2), são selecionadas as 3 páginas mais relevantes, desde que a diferença de pontuação entre elas não ultrapasse um valor limiar

definido. Sendo assim, se a página com maior pontuação não for classificada como relevante na Classificação Baseada em Gênero, ainda há outras duas possibilidades a serem analisadas.

A classificação baseada em gênero e conteúdo consiste em obter a interseção dos resultados da classificação baseada em gênero e classificação baseada em conteúdo. Caso haja mais de uma página relevante, é considerada a que tiver maior pontuação no *ranking* da classificação baseada em conteúdo. Em caso de empate, será considerada relevante, a que tiver melhor posicionamento no resultado da máquina de busca, uma vez que esta também possui estratégias que buscam apresentar as páginas mais relevantes nas primeiras posições.

Ao final desta etapa, tem-se as páginas devidamente classificadas como sendo ou não relevantes. O algoritmo seleciona a página relevante para cada ano de uma conferência, se houver alguma.

3.5 Extração de informação das páginas relevantes

Esta etapa recebe como entrada um conjunto de páginas relevantes, consideradas como sendo as páginas das conferências. O objetivo, nesta etapa, é analisar individualmente cada página da conferência e extrair seus dados. A saída desta etapa é um conjunto de dados extraídos para cada página.

Foi desenvolvido um algoritmo para extrair as seguintes informações: número da edição da conferência, título utilizado oficialmente naquela edição, sigla e data de ocorrência.

Para isso, o algoritmo utiliza-se de expressões regulares para procurar por padrões dentro da página HTML que permitem a extração de

informações. Como as páginas possuem formatos semelhantes, é possível obter tais informações de muitas delas.

Um dos padrões mais comuns é formado pelo número da edição, seguido do título do veículo de publicação, sigla e ano; por exemplo, “1st International Conference on Very Large Data Bases (VLDB 2015)”. Para extrair os dados neste padrão, utilizou-se a expressão regular da Figura 9.

```
.*?(?:(\\d{1,3})(?:\\<sup\\>)?(?:st|nd|rd|th)(?:\\<\\/sup\\>)?|
(\\d{4}))\\s([a-zA-Z\\s\\-]*)\\p{Punct}?(\\p{Upper}*)?[\\s*']
(\\d{4}|\\d{2})?.*
```

Figura 9 Expressão regular que extrai a edição, ano, título e sigla da conferência

Para as datas, procurou-se por padrões como: “August 31 - September 4, 2015”. Para extrair os dados com padrões parecidos a este, relacionados a data do evento, utilizou-se a expressão regular da Figura 10.

```
.*(January|February|March|April|May|June|July|August|September|October|
November|December)\\s(\\d{1,2})\\s*(?:\\<sup\\>)?(?:st|nd|rd|th)?
(?:\\<\\/sup\\>)?\\s?(to|-)\\s?(January|February|March|April|May|June|
July|August|September|October|November|December)*\\s?(\\d{1,2})\\s*
(?:\\<sup\\>)?(?:st|nd|rd|th)?(?:\\<\\/sup\\>)?,?\\s(\\d{4}|\\d{2})?.*
```

Figura 10 Expressão regular que extrai o dia, mês e ano da edição da conferência

3.6 Extração de informações do conjunto de dados

Esta etapa recebe um conjunto de dados para cada conferência, compostos por todos os dados obtidos nas etapas anteriores. Objetivou-se, nesta etapa, analisar o conjunto de dados geral de cada conferência e extrair informações não apenas de uma edição, mas do conjunto de suas edições. A saída desta etapa é um conjunto de informações gerais de cada conferência, que complementam os dados extraídos individualmente de cada página na etapa descrita na seção anterior.

Analisando-se os dados do conjunto de edições de cada conferência é possível, por exemplo, obter informações sobre em quais anos uma conferência foi realizada, quando ocorreu uma troca de nome, qual edição foi realizada em qual ano, no intuito de traçar um histórico de realização de cada conferência.

3.7 Exemplo da abordagem proposta

Considere o exemplo de uma conferência hipotética chamada Conferência UFLA (Universidade Federal de Lavras). Suponha que ela tenha surgido no ano de 1990, quando a UFLA ainda se chamava ESAL (Escola Superior de Agricultura de Lavras) e ocorreu anualmente até o ano 2015, exceto em 1997 que, por algum motivo, não ocorreu tal edição da conferência.

Na Figura 11, apresenta-se um exemplo de como seria o comportamento da abordagem proposta. A abordagem começa a coleta utilizando a lista de *strings* L para o ano mais recente (2015) e vai decrementando ano a ano e, para cada um, realiza consultas à máquina de busca, coletando, classificando as páginas e extraindo seu dados.

Considere que nas edições hachuradas, o classificador não identificou nenhuma página relevante. Conforme foi dito, em 1997 realmente não houve a realização da conferência, mas nota-se que em 2009 e 1998 o algoritmo não classificou uma página relevante. Isso pode ocorrer, em razão da indisponibilidade da página na Web ou a erros de classificação, que eventualmente podem ocorrer.

De 1993 até 1989, a abordagem novamente não classificou nenhuma página relevante. Depois de 5 anos, ele identifica uma possível troca de

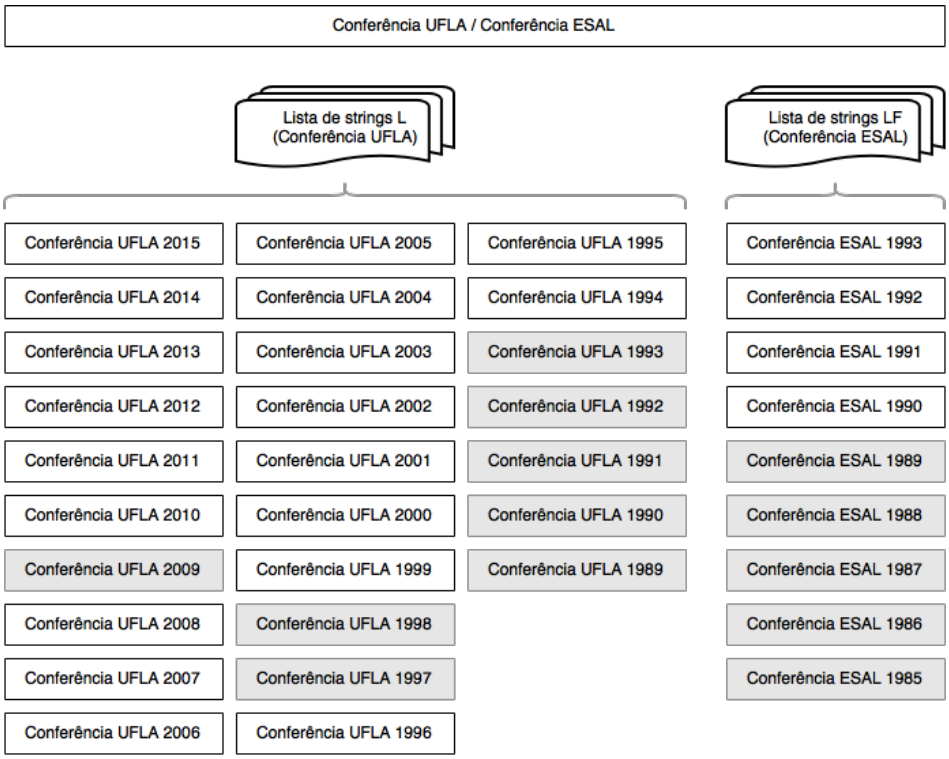

Figura 11 Exemplo da abordagem proposta aplicada à Conferência UFLA/ESAL

nome do veículo de publicação e passa a utilizar a lista *LF* para os títulos antigos. Após encontrar algumas edições da conferência com o nome antigo, de 1989 a 1985 a abordagem não encontra mais páginas relevantes e encerra a coleta.

Para cada edição, a abordagem faz a extração de dados na página relevante, exceto nas edições hachuradas, onde não foi encontrada página relevante. Além da extração de dados individuais das edições, a abordagem também faz uma extração de informações do conjunto de edições, obtendo dados como o registro histórico da conferência com ano de realização mais recente, ano de realização mais antigo, ano de mudança de nome. No exem-

plo, a “Conferência ESAL” começou em 1990, mudou de nome em 1994 para “Conferência UFLA” e continua sendo realizada atualmente.

Também é possível inferir informações como o número da edição. Considere, por exemplo, que foi extraído das edições da Conferência UFLA o número de cada edição. Em 2009, por algum motivo, o algoritmo não encontrou uma página relevante para extrair dados, mas observando-se que em 2008 aconteceu a 18ª edição e em 2010 a 20ª edição da conferência, pode-se inferir que em 2009 aconteceu a 19ª edição.

4 AVALIAÇÃO EXPERIMENTAL

Nesta seção, são descritos as bases de dados, métricas utilizadas, experimentos, resultados e discussões e os problemas, dificuldades e limitações.

4.1 Bases de dados

A partir do arquivo de autoridade criado por Pereira, Silva e Esmin (2014), foram obtidos siglas e títulos atuais e antigos de conferências. Com estes, foram geradas duas bases de dados denominadas C100 e CF107, nas quais os experimentos foram executados. As bases são formadas pelo conjunto de dados gerados pela abordagem. Os dados foram coletados entre janeiro e abril de 2015.

A base de dados C100 é formada por 100 conferências que possuem edição no ano de 2014. As coletas foram realizadas apenas das edições de 2014 das conferências. Elas foram selecionadas aleatoriamente do arquivo de autoridade. Para essas conferências, coletaram-se informações dos resultados da máquina de busca e das páginas HTML. Todas as páginas HTML foram classificadas manualmente por um especialista humano.

A base de dados CF107 é formada por 107 conferências que possuem, no arquivo de autoridade, informações de siglas e títulos antigos. Para esta base de dados, a coleta foi bem maior que a outra base, pois, neste caso, não se coletou apenas um ano para cada conferência, e sim vários anos, conforme a abordagem proposta constatasse a necessidade. O número de edições coletadas é particular para cada conferência.

4.2 Métricas de avaliação

Para avaliar os classificadores propostos neste trabalho, foram utilizadas as métricas acurácia, precisão, revocação e F1. A acurácia é a fração das páginas atribuídas a suas classes corretas pelo classificador. A acurácia, a precisão e a revocação são dadas por:

$$Acurácia = \frac{|a|}{|t|}$$

$$Precisão = \frac{|R \bigcap A|}{|A|}$$

$$Revocação = \frac{|R \bigcap A|}{|R|}$$

onde $|a|$ é o número de acertos na classificação de documentos, $|t|$ é o número total de documentos, $|A|$ é o número de páginas classificadas como relevantes pelo classificador, $|R|$ é o número de páginas classificadas como relevantes pelo especialista e $|R \bigcap A|$ é o número de páginas classificadas como relevante pelo classificador e que realmente são relevantes. A medida F1 é dada por:

$$F_1 = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação}$$

4.3 Experimentos, resultados e discussões

Os experimentos foram executados sobre as bases de dados C100 e CF107. Os experimentos sobre a base C100 permitiram a avaliação de diferentes estratégias no processo de classificação. Para cada conferência, foram submetidas consultas para coleta de páginas referentes apenas às edições do ano de 2014.

Na base CF107, com base nos resultados dos experimentos anteriores, foi possível avaliar a execução da abordagem completa. Uma amostra de dados da base CF107 foi classificada manualmente e utilizada em mais um conjunto de experimentos que visaram à exploração e avaliação de outras estratégias, em busca de melhores resultados.

4.3.1 Experimentos na base de dados C100

Foram executados 14 experimentos utilizando a base de dados C100, com diversas variações, com o objetivo de verificar uma maneira de obter melhores resultados nas classificações.

Utilizou-se a biblioteca LibSVM[16] para a classificação baseada em gênero. Para o treinamento e teste, as páginas de todas as 100 conferências coletadas foram unidas em uma única coleção e esta foi dividida aleatoriamente em 10 partes para uma validação cruzada (*cross-validation*), onde, em cada iteração, o SVM utiliza 9 partes para treinamento e 1 parte para teste. Antes da execução dos treinamentos, foi executado o grid-svm, que ajusta os melhores parâmetros para o SVM com os dados de entrada. Ao final, todas as partes foram classificadas pelo SVM e foi calculada a média aritmética para as métricas de avaliação.

Por meio dessa coleção, também foi criada uma lista de termos gerados automaticamente. Para isso, utilizou-se um algoritmo de seleção de *features* por meio dos seguintes passos e utilizando-se as seguintes configurações: os termos de todas as páginas classificadas manualmente foram importados no Weka e aplicou-se o filtro *weka.filters.unsupervised.attribute.StringToWordVector* utilizando-se o TF-

[16]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

IDF. Esse filtro converte atributos *string* em um conjunto de atributos que representam informações sobre a ocorrência da palavra a partir do texto contido nas *strings*. Para a geração automática dos termos de gênero, aplicou-se o filtro *weka.filters.supervised.attribute.AttributeSelection*, que é um filtro de atributos supervisionado bem flexível e que permite o uso de vários métodos de pesquisa e avaliação que podem ser combinados. O *evaluator* determina como os atributos/subconjuntos são avaliados. Nesse caso, utilizou-se o *weka.attributeSelection.InfoGainAttributeEval*, que avalia o valor de um atributo medindo o ganho de informação no que diz respeito à classe. O *search* determina o método de pesquisa e utilizou-se o *weka.attributeSelection.Ranker*, que cria um *ranking* dos atributos de acordo com suas avaliações individuais.

Quanto à classificação por conteúdo, os experimentos 1 a 7 utilizaram a EC1, que seleciona apenas uma página relevante, e os experimentos 8 a 14, a EC2, que seleciona até 3 páginas relevantes.

Essa série de experimentos foi realizada buscando-se analisar diferentes configurações baseadas em hipóteses criadas no decorrer da execução dos experimentos. Sendo assim, diversas configurações foram testadas e seus resultados coletados e analisados posteriormente, para que se encontrar alguma configuração que gere bons resultados.

a) Experimento 1: para a etapa de treinamento do SVM, foram utilizados 61 termos de gênero gerados manualmente por um especialista humano (Tabela 2 - coluna A). Esse número de termos é arbitrário, pois eles simplesmente foram sendo gerados com a análise das páginas coletadas e obtidos conforme constatou-se a necessidade. Ao final da análise obteve-se esse número de termos.

b) Experimento 2: fez-se uma redução dos termos gerados manualmente para 52, onde os que tinham o mesmo radical foram considerados como o mesmo termo (Tabela 2 - coluna B). O mesmo aplicou-se aos termos das páginas. Na Figura 12, apresenta-se um exemplo de duas ocorrências desse caso.

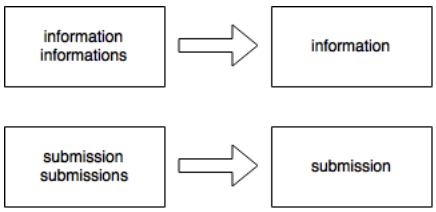

Figura 12 Exemplo da redução de termos ocorrida no Experimento 2

c) Experimento 3: foram usados os 100 primeiros termos de gênero gerados automaticamente (Tabela 3 - coluna A). O número de termos foi definido empiricamente. Os experimentos realizados com termos gerados automaticamente permitiram verificar se realmente é necessário que um humano faça um trabalho manual ou se um algoritmo de seleção de *features* é capaz de obter termos que gerem bons resultados, comparando-se com os resultados obtidos pelos termos gerados manualmente.

d) Experimento 4: fez-se uma redução dos termos, onde os que tinham o mesmo radical foram considerados como o mesmo termo, semelhantemente ao realizado no Experimento 2, porém aplicado aos termos gerados automaticamente. Com isso, o número de termos de gênero foi reduzido a 90 (Tabela 3 - coluna B).

Tabela 2 Listas de termos gerados manualmente por um especialista humano onde (A) é o conjunto original de termos gerados manualmente e (B) é o conjunto de termos reduzidos, onde foram removidos os termos com mesmo radical

A			B	
about	local	support	about	participant
abstract	location	techinical	abstract	partner
accepted	notification	tourism	accepted	past
accommodations	organization	travel	accommodation	poster
author	organizers	tutorial	author	presentation
awards	paper	university	award	proceeding
call	papers	venue	call	program
calls	participants	visa	chair	project
cfp	partners	workshops	comittee	registration
chair	past		conference	research
comittee	posters		date	scientific
comittees	presentation		deadline	scope
conference	proceedings		edition	slide
dates	program		event	speaker
deadline	projects		full	sponsor
editions	registration		important	steering
events	research		info	student
for	scientific		information	submission
full	scope		international	support
important	slides		invited	techinical
info	speakers		keynote	tourism
information	sponsors		loca	travel
informations	steering		notification	tutorial
international	students		organization	university
invited	submission		organizer	venue
keynote	submissions		paper	workshop

Tabela 3 Listas de termos gerados automaticamente onde (A) é o conjunto inicial de termos gerados pelo algoritmo de seleção de *features* e (B) é o conjunto de termos reduzidos, onde foram removidos os termos com mesmo radical

A			B		
registration	held	conference	acceptance	held	sign
dates	policy	projects	accepted	help	speaker
venue	demos	together	accommodation	home	sponsor
submission	instructions	materials	account	hotel	study
will	algorithm	browse	add	important	submission
important	article	supporters	alert	instruction	submit
call	pp	and	algorithm	invite	support
sponsors	submit	association	and	is	teaching
program	submitted	vol	approach	keynote	template
papers	paper	visa	are	latest	terms
terms	hotel	contacts	article	material	that
speakers	deadline	chapter	association	message	together
ready	resources	approach	author	my	travel
accommodation	keynotes	books	be	notification	tutorials
notification	ebooks	templates	books	ny	venue
accepted	add	panels	browse	panel	view
keynote	practitioners	extended	call	paper	visa
publications	tutorials	latest	camera	personal	vol
submissions	alerts	we	chair	policy	we
committee	sign	speaker	chapter	poster	welcome
welcome	is	articles	code	pp	will
be	feedback	study	committee	practitioner	workshop
my	search	eng	conditions	print	
travel	contents	invite	conference	privacy	
workshops	print	acceptance	contact	professional	
camera	sponsored	education	content	program	
conditions	that	account	date	project	
researchers	ny	professional	deadline	publications	
committees	view	message	demos	ready	
invited	personal	posters	ebooks	registration	
sessions	poster	teaching	education	researchers	
privacy	chairs	help	eng	resources	
authors	code		extended	search	
are	home		feedback	session	

e) Experimentos 5 e 6: uma página HTML é composta por *tags* que marcam elementos específicos e permitem organizá-la e estruturá-la. Os menus, por exemplo, fazem a ligação da página atual com outras páginas por meio de *links*, que utilizam a *tag* <a>. Considerando o fato de que os termos usados nos menus das páginas de conferência possuem uma certa semelhança, optou-se por realizar experimentos em que fossem considerados como termos das páginas todos aqueles que estivessem apenas entre as *tags* <a> e , ao invés de todas as palavras que aparecem na página. Os Experimentos 5 e 6 foram executados com essas configurações, utilizando os 61 termos de gênero gerados manualmente e com os 100 termos gerados automaticamente, respectivamente.

f) Experimento 7: possui as mesmas configurações do Experimento 1, exceto por ter sido removido, na classificação baseada em conteúdo, o filtro que garante ao menos uma menção à conferência e uma ao ano.

g) Experimento 8: utiliza as mesmas configurações do Experimento 3, exceto por utilizar a estratégia EC2 na classificação baseada em conteúdo.

h) Experimento 9: utiliza as mesmas configurações do Experimento 8, porém alterou-se os termos de gênero, onde foi feita uma redução dos termos com a remoção manual de *stopwords* e verbos comuns utilizados em páginas que não são de conferências, totalizando 52 termos. Esse número de termos é arbitrário, uma vez que foi obedecida uma regra que filtrou o conjunto de termos, continuando apenas os termos permitidos pela regra.

i) Experimento 10: reduziu-se ainda mais o número de termos do Experimento 9, deixando apenas 35 termos diretamente ligados a conferências. A quantidade de termos resultantes também é arbitrária.

j) Experimento 11: o conjunto de termos do experimento anterior foi acrescido de termos relevantes seguindo a lista de termos gerados automaticamente, aumentando a lista novamente até completar 80 termos.

k) Experimentos 12, 13 e 14: utilizando as configurações do Experimento 8, aplicaram-se estratégias de pós-processamento entre as etapas de classificação baseada em gênero e classificação baseada em conteúdo. A ideia é baseada na intuição de que nessas tags, que são links, ficaria mais fácil de se encontrar alguns tipos de termos, como os que comumente são encontrados nos menus das conferências e, assim, melhorar a classificação das páginas. No Experimento 12, o algoritmo verifica o conteúdo entre as *tags* <a> e das páginas em busca de termos comuns nos menus de páginas. Caso a página classificada como relevante na classificação baseada em gênero não atinja um determinado limiar nesta verificação, ela passa a ser considerada como não-relevante. No Experimento 13, o algoritmo faz a verificação de forma inversa ao Experimento 12. Ele faz a verificação nas páginas classificadas como não-relevantes e, caso elas ultrapassem o valor do limiar, elas passam a ser consideradas como relevantes. O Experimento 14 faz uma verificação em busca de sentenças como "*call for papers*"e "*important dates*", por exemplo. Caso a página classificada como relevante não possua 50% do conjunto de sentenças, ela passa a ser considerada como não-relevante.

Tabela 4 Resultados da Classificação Baseada em Gênero (CBG) para os Experimentos 1 a 14

Experimento	Acurácia (%)	P (%)	R (%)	F1 (%)
1	$89,4 \pm 1,9$	$86,2 \pm 5,1$	$77,4 \pm 4,7$	$81,5 \pm 3,5$
2	$89,8 \pm 1,6$	$85,9 \pm 2,9$	$78,6 \pm 5,3$	$82,1 \pm 3,4$
3	$91,6 \pm 1,0$	$88,4 \pm 4,0$	$83,3 \pm 3,3$	$85,8 \pm 1,7$
4	$89,0 \pm 2,7$	$84,6 \pm 4,3$	$78,6 \pm 4,7$	$81,4 \pm 3,7$
5	$89,7 \pm 1,4$	$89,8 \pm 3,2$	$74,0 \pm 4,9$	$81,1 \pm 2,8$
6	$90,2 \pm 1,0$	$87,9 \pm 4,2$	$77,5 \pm 2,5$	$82,4 \pm 2,2$
7	$89,4 \pm 5,9$	$86,2 \pm 5,1$	$77,4 \pm 4,7$	$81,5 \pm 3,5$
8	$91,6 \pm 1,0$	$88,4 \pm 4,0$	$83,3 \pm 3,3$	$85,8 \pm 1,7$
9	$90,0 \pm 1,2$	$85,4 \pm 2,5$	$79,7 \pm 3,4$	$82,5 \pm 2,5$
10	$90,4 \pm 1,0$	$85,6 \pm 3,8$	$80,9 \pm 2,9$	$83,1 \pm 2,3$
11	$90,4 \pm 1,3$	$85,2 \pm 3,5$	$93,2 \pm 3,3$	$84,2 \pm 1,4$
12	$91,6 \pm 1,0$	$88,4 \pm 4,0$	$83,3 \pm 3,3$	$85,8 \pm 1,7$
13	$91,6 \pm 1,0$	$88,4 \pm 4,0$	$83,3 \pm 3,3$	$85,8 \pm 1,7$
14	$91,6 \pm 1,0$	$88,4 \pm 4,0$	$83,3 \pm 3,3$	$85,8 \pm 1,7$

Nas Tabelas 4, 5 e 6, apresentam-se os resultados obtidos nos experimentos 1 a 14. Nas Tabelas 4 e 6, que trabalham com a média de resultados, há a informação do intervalo de confiança considerando-se 95% de confiança.

Tabela 5 Resultados da Classificação Baseada em Conteúdo (CBC) para os Experimentos 1 a 14

Experimento	Acurácia (%)	P (%)	R (%)	F1 (%)
1	87,0	88,8	87,0	87,9
2	87,0	88,8	87,0	87,9
3	87,0	88,8	87,0	87,9
4	87,0	88,8	87,0	87,9
5	87,0	88,8	87,0	87,9
6	87,0	88,8	87,0	87,9
7	88,0	88,0	88,0	88,0
8	-	-	-	-
9	-	-	-	-
10	-	-	-	-
11	-	-	-	-
12	-	-	-	-
13	-	-	-	-
14	-	-	-	-

Tabela 6 Resultados da Classificação Baseada em Gênero e Conteúdo (CBGC) para os Experimentos 1 a 14

Experimento	Acurácia (%)	P (%)	R (%)	F1 (%)
1	$77,0 \pm 1,9$	$98,8 \pm 5,1$	$77,0 \pm 4,7$	$86,5 \pm 3,5$
2	$76,0 \pm 1,6$	$96,2 \pm 2,9$	$76,0 \pm 5,3$	$84,9 \pm 3,4$
3	$80,0 \pm 1,0$	$97,6 \pm 4,0$	$80,0 \pm 3,3$	$87,9 \pm 1,7$
4	$75,0 \pm 2,7$	$96,1 \pm 4,3$	$75,0 \pm 4,7$	$84,3 \pm 3,7$
5	$75,0 \pm 1,4$	$98,7 \pm 3,2$	$75,0 \pm 4,9$	$85,2 \pm 2,8$
6	$80,0 \pm 1,0$	$98,8 \pm 4,2$	$80,0 \pm 2,5$	$88,4 \pm 2,2$
7	$80,0 \pm 5,9$	$97,6 \pm 5,1$	$80,0 \pm 4,7$	$87,9 \pm 3,5$
8	$83,0 \pm 1,0$	$93,3 \pm 4,0$	$83,0 \pm 3,3$	$87,8 \pm 1,7$
9	$80,0 \pm 1,2$	$90,9 \pm 2,5$	$80,0 \pm 3,4$	$85,1 \pm 2,5$
10	$81,0 \pm 1,0$	$93,1 \pm 3,8$	$81,0 \pm 2,9$	$86,6 \pm 2,3$
11	$83,0 \pm 1,3$	$91,2 \pm 3,5$	$83,0 \pm 3,3$	$86,9 \pm 1,4$
12	$76,0 \pm 1,0$	$91,6 \pm 4,0$	$76,0 \pm 3,3$	$83,1 \pm 1,7$
13	$76,0 \pm 1,0$	$91,6 \pm 4,0$	$76,0 \pm 3,3$	$83,1 \pm 1,7$
14	$83,0 \pm 1,0$	$93,3 \pm 4,0$	$83,0 \pm 3,3$	$87,8 \pm 1,7$

Nos resultados obtidos na classificação baseada em gênero, ao analisar o valor da acurácia, os experimentos 3, 8, 12, 13 e 14 obtiveram os maiores valores, com 91,6% na taxa de acertos, porém, eles ficaram estatisticamente empatados com os demais resultados. Os experimentos 8, 12, 13 e 14 são variações do Experimento 3, que não modificaram as configurações da classificação baseada em gênero; por isso, o resultado para todos eles permaneceram inalterados. Para as demais métricas, os experimentos também permaneceram estatisticamente empatados.

Comparando-se os termos gerados manual e automaticamente, percebe-se uma ligeira melhora nos resultados usando-se os gerados automaticamente. Assim, não é necessário que um especialista humano se preocupe em gerar uma lista de termos de gênero, uma vez que existem algoritmos de seleção de *features* capazes de realizar essa tarefa com resultados tão bons quanto os do especialista humano.

Analisando-se a classificação baseada em conteúdo, os resultados foram os mesmos, pois todos utilizaram a mesma estratégia, exceto o 7, onde foi removido o filtro que garante que a página classificada faça menção obrigatoriamente ao ano e ao título ou sigla do veículo de publicação. Embora o resultado tenha melhorado um pouco, o algoritmo está confiando no resultado da máquina de busca, ao invés de fazer sua própria classificação; por isso, não é uma boa estratégia para ser aplicada neste contexto. Os experimentos 8 a 14 não possuem resultados na classificação baseada em conteúdo, pois não há a classificação de uma única página; só faz sentido se avaliar a classificação baseada em gênero e conteúdo.

De modo geral, a acurácia e revocação da classificação baseada em gênero e conteúdo foram inferiores aos da baseada apenas em conteúdo. Isso

ocorreu, em razão de o fato de muitas páginas classificadas como relevantes na classificação baseada em conteúdo não terem sido classificadas como relevantes pela classificação baseada em gênero.

Estatisticamente, os melhores resultados para a classificação baseada em conteúdo e a classificação baseada em gênero e conteúdo estão empatadas. Isso indica que a classificação baseada em conteúdo é mais recomendada para essa aplicação, uma vez que ela é mais simples e eficiente de se implementar.

4.3.2 Experimentos na base de dados CF107

Objetivou-se, nos próximos experimentos, avaliar a coleta dos dados de cada edição de cada conferência, de forma a obter suas informações históricas. Foi utilizada a base de dados CF107, formada por 107 conferências, as quais tiveram mudanças de nome ao longo de sua história.

O Experimento 15 foi realizado sobre uma amostra aleatória, classificada manualmente, composta por 6 das conferências da base de dados, utilizando-se as mesmas configurações do Experimento 6, o qual teve os maiores valores na classificação baseada em gênero e conteúdo. Foram coletados dados de 153 possíveis edições dessas conferências. Na Tabela 7, apresenta-se os resultados obtidos. Como a CBG e a CBGC trabalham com a média, a tabela exibe também seu intervalo de confiança, considerando-se a confiança de 95%.

Tabela 7 Resultados do Experimento 15

	Acurácia (%)	Precisão (%)	Revocação (%)	F1 (%)
CBG	$89,8 \pm 1,0$	$94,2 \pm 4,2$	$53,1 \pm 2,5$	$67,9 \pm 2,2$
CBC	90,2	90,2	68,7	78,0
CBGC	$83,3 \pm 1,0$	$83,3 \pm 4,2$	$63,5 \pm 2,5$	$72,12 \pm 2,2$

Uma boa precisão é importante para que a coleta de dados seja consistente, enquanto uma boa revocação permite que o algoritmo recupere mais páginas relevantes e assim, permitindo que as edições anteriores possam ser recuperadas. Com isso, a medida F1 mostra que a classificação baseada em conteúdo obteve os melhores resultados para o experimento executado.

Os Experimentos 16 e 17 foram executados sobre a mesma amostra de dados do Experimento 15. Utilizou-se o esquema de peso IHF (*Inverse Host Frequency*) (TAN; KAN; LEE, 2006). A ideia é que utilizando-se esse recurso, as páginas cujo *host* sejam muito frequentes não sejam classificadas como relevantes, com isso eliminando páginas de bibliotecas digitais e sites de divulgação de eventos.

Semelhante à medida *Inverse Document Frequency*, utilizada na recuperação de informação, Tan, Kan e Lee (2006) formularam a medida IHF para avaliar a raridade de um *host* da Internet entre um conjunto de documentos da Web.

Dado o conjunto de URLs obtido por meio de consultas a uma máquina de busca, cada URL é truncada em seu *hostname*. Se um *hostname* h possui frequência $f(h)$, então seu IHF é calculado como:

$$IHF(h) = log_2 \frac{max_h f(h) + 1}{f(h) + 1} + 1$$

Considerando o exemplo das URLs "http://www.icme2015.ieee-icme.org/callforpapers.php" e "http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=32420", os *hostnames* obtidos seriam "www.icme2015.ieee-icme.org" e "www.ieee.org", respectivamente.

O objetivo do uso do IHF nos experimentos é reduzir o número de páginas de bibliotecas digitais e sites de divulgação de eventos que foram classificadas como relevantes. Dessa forma, *hostnames* como "dl.acm.org", "www.ieee.org" e "www.wikicfp.com", por exemplo, obteriam IHF mais baixo, em relação a *hostnames* como "www.icme2015.ieee-icme.org" e "ase2013.org", por exemplo.

O valor do IHF é baixo para os *hosts* mais frequentes, enquanto é mais alto para os *hosts* mais raros, como pode ser visto nestes dois exemplos, obtidos nesta base de dados, onde $IHF(www.icme2015.ieee\text{-}icme.org) = 4,1$ e $IHF(dl.acm.org) = 1,0$.

No Experimento 16, foi definido um valor limiar e as páginas que tiveram o IHF menor que esse valor foram eliminadas. Esse limiar foi obtido empiricamente, analisando-se os valores do IHF obtidos nas páginas de bibliotecas digitais e sites de divulgação de eventos e comparando-os com o valor do IHF de páginas de conferências. No Experimento 17, o valor do IHF para cada página foi incorporado ao valor do *ranking* calculado na classificação baseada em conteúdo. O valor obtido no IHF foi multiplicado por 0,10 e somado ao valor do *ranking*, dessa maneira, as páginas cujos *hosts* possuem o valor do IHF mais alto recebem um valor maior a ser somado que as páginas cujos *hosts* possuem o valor do IHF mais baixo. Na Tabela 8, apresentam-se os resultados obtidos para os experimentos envolvendo o IHF. Em razão da característica da classificação baseada em conteúdo, em que as métricas são calculadas para cada edição ou ano e não para cada documento, como ocorre na classificação baseada em gênero, a acurácia é igual a precisão, por isso não foi apresentada aqui.

Tabela 8 Resultados dos Experimentos 16 e 17

Experimento	Precisão (%)	Revocação (%)	F1 (%)
16	78,9	47,6	59,4
17	81,3	63,5	71,3

Os resultados obtidos nesses experimentos foram inferiores aos obtidos no Experimento 15, apresentados na Tabela 7.

A estratégia do IHF permite a remoção de *hosts* com maior frequência, eliminando-se assim as páginas de bibliotecas digitais, por exemplo. Porém, notou-se que algumas páginas de conferências utilizam *hosts* comuns (Ex: "sites.google.com"), o que pode ter provocado uma queda nos resultados. Por isso, optou-se por não utilizar essa estratégia.

O Experimento 18 foi realizado sobre a base de dados completa, utilizando-se a CBC. Cerca de 2.380 possíveis edições foram analisadas, totalizando mais de 32.400 páginas coletadas, cujos resultados são apresentados na Tabela 9. A revocação não foi calculada em decorrência do elevado número de páginas a serem classificadas manualmente.

Tabela 9 Resultados do Experimento 18

	Páginas Relevantes	Acertos	Precisão (%)
Siglas e Títulos Atuais	740	553	74,7
Siglas e Títulos Antigos	41	21	51,2
TOTAL	781	574	73,5

O número de erros (207) foi influenciado pelo fato de que, para 161 páginas, a máquina de busca não recuperou a página oficial da conferência e o algoritmo acabou classificando como relevante alguma outra página que continha dados sobre a mesma. Para outras 46 páginas, o algoritmo realmente errou a classificação das páginas relevantes.

Do total de páginas classificadas como relevantes, o algoritmo conseguiu extrair o número da edição de 142 páginas e, em outras 381, ele identificou que não havia o número da edição na frente do título da conferência, e sim o ano. Foram extraídas corretamente 477 siglas em várias edições de conferências, o mês de realização de 344 e os dias de 343 edições.

Na extração de siglas, foram 140 erros causados por diversos motivos como: uso de padrões fora dos definidos na expressão regular, por haver siglas de mais de uma conferência na mesma página (conferências que ocorrem simultaneamente) e o extrator coletar a sigla de outra conferência; e páginas classificadas incorretamente, fazendo com que o extrator extraia algo que não seja relacionado à conferência. Na extração dos dias e meses de realização, ocorreram 54 e 55 erros, respectivamente. A maioria causados pela extração de outras datas contidas nas páginas, como o prazo de submissão de artigos, por exemplo.

Em 38 páginas, o processo de extração de dados foi prejudicado pelo fato de as informações estarem dispostas em imagens, e o extrator, sendo baseado em dados de texto, não conseguiu efetuar as extrações. Várias páginas de conferências não trazem o número da edição e sim o ano; por isso, o extrator não conseguiu obter tal informação. Em alguns casos, também ocorreu de o extrator encontrar outro tipo de informação com o mesmo padrão buscado e extraí-la.

A abordagem foi capaz de coletar um histórico maior do que 5 edições para 74 conferências, além de identificar a mudança de nome de 21 delas, e 81 tiveram uma edição realizada no ano 2014.

Em alguns casos, o baixo número de edições de uma conferência se deu ao fato de ela não possuir as páginas oficiais disponíveis. Muitas delas possuem informações apenas em sites de bibliotecas digitais.

As conferências com maior número de edições identificadas geralmente possuem domínios (URLs) próprios para suas edições, com uma gestão centralizada, tornando disponíveis as páginas de edições antigas. Já as conferências com menor número de edições identificadas possuem domínios diferentes para cada edição. Assim, após algum tempo, elas acabam ficando indisponíveis, permanecendo registros de tais edições apenas em bibliotecas digitais ou *sites* de terceiros. Nesses casos, a abordagem realmente não identificou a página por estar indisponível ou identificou incorretamente uma com características muito semelhantes a uma página de conferência.

Nas consultas envolvendo o título antigo da conferência, ocorreram 44 casos em que as páginas existem, mas a máquina de busca não as recuperou. Em outros casos, as páginas realmente não foram encontradas por meio de diversas consultas e buscas manuais.

4.4 Dificuldades, problemas e limitações

Classificar páginas Web e extrair informações delas não são tarefas triviais, principalmente pelo fato de elas não terem uma estrutura padronizada. Como as páginas das conferências que ocorrem no mundo todo são construídas por seus organizadores, não existe um padrão específico de disposição dos dados.

Uma limitação encontrada foi no uso da *Google Search Engine API*, que limita a quantidade de consultas que podem ser realizadas de forma gratuita, o que estende o tempo para coleta das informações.

Um dos problemas encontrados na classificação é a ocorrência de páginas muito parecidas que possuem informações da conferência, mas não são a sua página oficial. Geralmente são *sites* de divulgação de eventos científicos e afins, e que não são interessantes para o propósito deste trabalho, pois podem conter erros na replicação dos dados da conferência.

Outra limitação ocorre quando a página da conferência é removida ou o servidor em que a página está hospedada está *offline* no momento da coleta. Isso ocorreu principalmente nas consultas por títulos antigos.

Em páginas mais antigas, é comum que a disposição de informações do cabeçalho da página e de seus menus estejam em imagens, o que dificultou o processo de classificação, já que esta abordagem trabalha exclusivamente com textos.

5 CONCLUSÃO E TRABALHOS FUTUROS

Coletar, classificar e extrair informações de páginas de veículos de publicação não são tarefas triviais, pois não há uma padronização de seus conteúdos.

Os resultados deste trabalho demonstram que é possível, por meio da submissão de consultas a uma máquina de busca, obter páginas de veículos de publicações com informações que podem ser extraídas (número da edição, ano da edição, sigla, título e data de realização do evento), e assim, enriquecer um arquivo de autoridade.

Os resultados obtidos de uma máquina de busca nem sempre são consistentes com o que se deseja; por isso, é necessária uma classificação mais precisa dos mesmos. Este trabalho abordou duas estratégias de classificação aplicadas ao problema de classificação de páginas relevantes de veículos de publicação. A classificação baseada apenas em conteúdo obteve os melhores resultados na execução da abordagem proposta.

Um extrator simples, baseado em expressões regulares, foi capaz de coletar informações como número da edição, ano, título da conferência, sigla e datas de realização.

Foi feita também uma avaliação dos problemas encontrados, e o principal deles é que as edições mais antigas das conferências não possuem mais páginas disponíveis na Web.

Como trabalhos futuros, a abordagem proposta para conferências será adaptada para periódicos e workshops. Geralmente, conferências e workshops possuem páginas diferentes para cada edição e é comum workshops ocorrerem junto com conferências. Já os periódicos, normalmente possuem uma página única, dentro da qual se encontram as edições.

Adaptações na abordagem proposta poderão permitir que se faça a coleta e extração de dados também desses veículos de publicação, ajudando a enriquecer ainda mais o arquivo de autoridade.

A utilização de outras estratégias de classificação também será realizada como trabalho futuro. Por exemplo, a substituição do classificador baseado no SVM por um classificador baseado na similaridade, permitindo o uso de ponderação para os termos de gênero e de conteúdo.

Outra estratégia a ser adotada em trabalhos futuros é a obtenção dos termos de gênero por meio da interseção dos termos de páginas relevantes. Nesse caso, um especialista humano classifica manualmente um conjunto de páginas, obtém-se os termos de cada página relevante e um algoritmo obtém um conjunto de termos mais comuns no conjunto de páginas.

Na etapa de coleta das páginas, caso a página não esteja disponível no momento, pode-se utilizar o conteúdo do cache da máquina de busca. Dessa forma, mesmo que a página não esteja disponível, será possível classificá-la e extrair seus dados normalmente.

E ainda, o extrator será estendido para coletar informações adicionais, tais como o local de realização e a lista dos artigos publicados. Além disso, pode-se utilizar ferramentas específicas para extrair dados das imagens, uma vez que a abordagem proposta trabalha apenas com dados textuais.

REFERÊNCIAS

ABBURU, S.; BABU, G. S. A framework for web information extraction and analysis. **International Journal of Computers and Technology**, Vin Rose Way, v. 7, n. 2, p. 574-579, June 2013.

ALFRED, R. et al. A robust framework for web information extraction and retrieval. **International Journal of Machine Learning and Computing**, Jurong West, v. 4, n. 2, p. 146-150, 2014.

ASSIS, G. de et al. A genre-aware approach to focused crawling. **World Wide Web**, New York, v. 12, n. 3, p. 285-319, 2009.

ASSIS, G. T. et al. The impact of term selection in genre-aware focused crawling. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2008, New York. **Proceedings…** New York: ACM, 2008. p. 1158-1163.

ASSIS, G. T.; LAENDER, A. H. F.; GONÇALVES, M. A. **Uma abordagem baseada em gênero para coleta temática de páginas da Web**. 2008. 60 p. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

AULD, L. Authorithy control: an eight-year review. **Library Resources & Technical Services**, Chicago, v. 26, n. 4, p. 319-330, 1982.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval:** the concepts and technology behind search. Indianapolis: Addison-Wesley Professional, 2011. 913 p.

CONNAWAY, L. S.; DICKEY, T. J. Publisher names in bibliographic data. **Library Resources and Technical Services Journal**, Chicago, v. 55, n. 4, p. 182-194, 2011.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Boston, v. 20, n. 3, p. 273-297, 1995.

COWIE, J.; LEHNERT, W. Information extraction. **Communications ACM**, New York, v. 39, n. 1, p. 80-91, Jan. 1996.

ETZIONI, O. et al. Open information extraction from the web. **Communications ACM**, New York, v. 51, n. 12, p. 68-74, Dec. 2008.

FERREIRA, A. A.; GONÇALVES, M. A.; LAENDER, A. H. A brief survey of automatic methods for author name disambiguation. **SIGMOD Rec**, New York, v. 41, n. 2, p. 15-26, Aug. 2012.

FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E. Using clustering strategies for creating authority files. **Journal of the American Society for Information Science**, Silver Spring, v. 51, n. 8, p. 774-786, 2000.

GUNASUNDARI, R.; KARTHIKEYAN, S. A new approach for web information extraction. **International Journal of Computer Technology and Applications**, Chennai, v. 3, n. 1, p. 211-215, 2012.

JALIL, K. A.; KAMARUDIN, M.; MASREK, M. Comparison of machine learning algorithms performance in detecting network intrusion. In: INTERNATIONAL CONFERENCE ON NETWORKING AND INFORMATION TECHNOLOGY, 2010, Manila. **Proceedings…** Manila: IEEE, 2010. p. 221-226.

KAMBER, J. H.; PEI, J. (Ed.). **Data mining**. 3rd ed. Boston: Morgan Kaufmann, 2012. (The Morgan Kaufmann Series in Data Management Systems). Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780123814791000162>. Acesso em: 10 fev. 2015.

KAVULYA, J. M. Digital libraries and development in sub-saharan africa: a review of challenges and strategies. **The Electronic Library**, Bingley, v. 25, n. 3, p. 299-315, 2007.

LEE, D. Practical maintenance of evolving metadata for digital preservation: algorithmic solution and system support. **International Journal on Digital Libraries**, Berlin, v. 6, n. 4, p. 313-326, July 2007.

LEE, D. et al. Are your citations clean? **Communications ACM**, New York, v. 50, n. 12, p. 33-38, Dec. 2007.

MA, Y.; CLEGG, W.; O'BRIEN, A. Digital library education: the current status. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 6., 2006, Chapel Hill. **Proceedings…** Chapel Hill: University of North Carolina, 2006. p. 165-174.

ONLINE COMPUTER LIBRARY CENTER. **VIAF:** the virtual international authority file. 2015. Disponível em: <https://viaf.org>. Acesso em: 10 abr. 2015.

PEREIRA, D. A. et al. Using web information for author name disambiguation. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 9., 2009, New York. **Proceedings…** New York: ACM, 2009. p. 49-58.

PEREIRA, D. A. et al. Using web information for creating publication venue authority files. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 8., 2008, New York. **Proceedings...** New York: ACM, 2008. p. 295-304.

PEREIRA, D. A.; SILVA, E. E. B. da; ESMIN, A. A. A. Disambiguating publication venue titles using association rules. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 14., 2014, Piscataway. **Proceedings…** Piscataway: IEEE, 2014. p. 77-85.

ROSA, I. B.; SHMORGUN, I.; LAMAS, D. Enabling mobile access to digital libraries in digital divide contexts. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES, 7., 2012, Madrid. **Proceedings…** Madrid: IEEE, 2012. p. 1-4.

SILVA, A. J. C. et al. Finding what is missing from a digital library: a case study in the computer science field. **Information Processing & Management**, Tarrytown, v. 45, n. 3, p. 380-391, May 2009.

TAN, Y. F.; KAN, M. Y.; LEE, D. Search engine driven author disambiguation. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 6., 2006, Chapel Hill. **Proceedings…** Chapel Hill: ACM, 2006. p. 314-315.

TUMER, D.; SHAH, M.; BITIRIM, Y. An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In: INTERNATIONAL CONFERENCE ON INTERNET MONITORING AND PROTECTION, 4., 2009, Venice. **Proceedings…** Venice: IEEE, 2009. p. 51-55.

VAPNIK, V. N. **The nature of statistical learning theory**. New York: Springer-Velag, 1995. 314 p.

WEINBERG, Z. **Svm separating hyperplanes (SVG).svg**. 2012. Disponível em: <https://commons.wikimedia.org/wiki/File%3ASvm_separating_hyperplanes_(SVG).svg>. Acesso em: 10 abr. 2015.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data mining:** practical machine learning tools and techniques. 3rd ed. Burlington: Morgan Kaufmann, 2011. 664 p. (The Morgan Kaufmann Series in Data Management Systems).