Journal of Computer Sciences

Original Research Paper

Enriching an Authority File of Scientific Conferences with Information Extracted from the Web

Heider Alvarenga de Jesus and Denilson Alves Pereira

Department of Computer Science, Universidade Federal de Lavras, PO Box 3037, 37.200-000, Lavras, Brazil

Article history
Received: 31-10-2016
Revised: 20-02-2017
Accepted: 05-05-2017

Corresponding Author:
Denilson Alves Pereira
Department of Computer Science, Universidade Federal de Lavras, PO Box 3037, 37.200-000, Lavras, Brazil
Email: denilson.pereira@gmail.com

Abstract: Authority files maintain variant forms to refer to the same entity and they are very useful in digital libraries. However, collect data and keep an updated authority file is not a trivial task. This paper proposes an approach for the enrichment of a publication venue authority file by extracting information on conferences from their web pages. Collecting additional data is important to improve the effectiveness of data disambiguation tools and information retrieval, such as those that measure the quality of a scientific publication based on bibliometrics (e.g., Journal Impact Factor). Most applications use only basic citation metadata, such as author's names, work and publication venue titles. However, data external to the publication, contained in the publication venue web page, can be very useful in the disambiguation task. Our approach includes the steps for querying a web search engine, classifying documents obtained in the result sets and extracting information from the relevant pages. We evaluated two methods for classifying documents, one based on genre and content and one based on content only. The experiments show good results to trace a history of conference editions, with data such as URL, year of each edition and dates of changing in their names.

Keywords: Authority File, Publication Venue, Web Search Engine, Classification, Information Extraction

Introduction

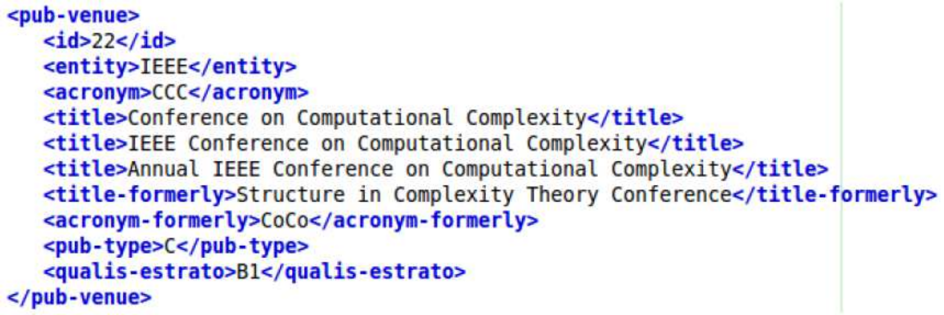
The quality of a scientific publication can be measured by bibliometrics such as Journal Impact Factor (JIF) (http://wokinfo.com/products_tools/analytical/jcr), SCImago Journal and Country Rank (SJR) (http://www.scimagojr.com) and Qualis Capes (https://sucupira.capes.gov.br) (Brazilian system). However, to make these measurements accurately, it is important to correctly identify the title (name) of the publication venues extracted from citations to scientific papers.

Bibliographic indexing systems of digital libraries work with a vast amount of data and constantly deal with inconsistent records. Citations to publication venues such as journals, conferences and workshops often have misspellings, variations in writing and abbreviations that make search and retrieval more difficult (Pereira *et al*., 2008).

A solution found by digital libraries to attempt to circumvent the problem of identifying variant forms to refer to the same entity is through the use of authority files. According to Auld (1982), an authority file maintains variations in writing used for a specific bibliographic attribute. However, the creation of authority files is not a simple task to be automated.

In the work Pereira *et al.* (2014), it was built a Computer Science publication venue authority file. This authority file consists of a set of records with information on current titles and acronyms, formerly titles and acronyms, among other information on publication venues. Figure 1 shows an example of a record in that authority file. From now on, we will refer to this authority file as Publication Venue Authority File (PVAF). A tool to search the PVAF is available at http://pvaf.dcc.ufla.br.

In addition to variations in name writing, authority files can also store other data related to entities, thus allowing better use in information retrieval. Information on conference edition numbers, years of occurrence, years of name changes (if occurred) and URLs, for example, does not exist in the authority file of Pereira *et al.* (2014) and they would be important for its enrichment.



```
<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>
```

Fig. 1. Example of a record in PVAF

The hypothesis in this study is that such information can be obtained from the Web by submitting queries to a search engine. The major challenges are to identify the relevant documents on the query result sets and extract the data contained therein, since there is no standard for Web page formats.

This paper presents an approach to collect web pages related to scientific conferences, identify their official pages and extract additional information about them in order to enrich the authority file created by Pereira *et al*. (2014). Due to the peculiarities of pages of different types of publication venues such as journals and conferences, the collecting strategies need to be adapted to each type of publication venue. In this study, we discuss strategies for collecting information on conferences.

The approach is composed of the following steps. First, it submits queries to a web search engine and through the result sets of these queries, it gets their URLs and gathering up their HTML pages. Then, the pages are classified, with the aim to select those relevant to data extraction. The official pages for each edition of the conferences are considered as relevant. One strategy used in the classification is an approach based on genre and content (Assis *et al*., 2008), where the classifier analyzes the terms referring to page structure (genre) and, subsequently, the terms relating to the publication venue itself (content). In another strategy, it was used only the classification based on content. Then, an extractor based on regular expressions scans the pages classified as relevant looking for the edition number, venue title, acronym and date of occurrence and finally, from the data extracted from each conference edition it draws up a history of its editions, including identification of dates of name changing when they exist.

Collecting these additional data is important to improve the effectiveness of data disambiguation tools and information retrieval (Sun *et al*., 2016). Websites of each conference edition, for example, are important to web archiving and for extracting additional information. PVAF was built from the elaboration of specific data extractors for some digital libraries such as ACM Digital Library and DBLP. However, these digital libraries do not have all this additional information nor they are available in a structured way. Another alternative would be the use of a focused crawler, however it is difficult to find a set of seed pages for this type of collection. Sites of digital libraries and specific conferences usually do not have links to other conference pages and call for paper sites are not complete. The collecting approach through a web search engine is suited to this problem, being generic and effective as demonstrated by our experiments.

In short, the main contributions of this work are (a) a proposal of an approach to collect, identify and extract information from official pages of conferences, (b) a set of experiments that demonstrate the effectiveness of the proposed approach and (c) the gathering of a set of data to enrich PVAF.

The remainder of this paper is organized as follows. Section 2 describes related works, section 3 details PVAF, section 4 describes our approach to obtain information on conferences from the Web, section 5 presents the experimental evaluation and, finally, section 6 presents the conclusions and future work.

Related Work

Maintain consistency of citations collected and stored in digital libraries is not a trivial task. Among the main problems are errors in data entry, many citation styles, lack of standards, imperfect collector software, ambiguous author names and abbreviations of publication venue titles (Lee *et al*., 2007).

Some digital libraries use authority files and work to maintain their consistency, thereby reducing problems such as the ambiguity of author names. The Virtual International Authority File (VIAF) (VIAF, 2016) is one of the major international initiatives in this direction. Creating an authority file is not a simple task. In French *et al*. (2000), the authors investigated the use of a number of approximate string matching techniques to create authority files for author affiliations. In Connaway and Dickey (2011), they worked with publisher name authority files.

Table 1. Summary of work related to our tasks

Task	Related work	Our work
Authority File creation and enrichment	(French *et al*., 2000; Pereira *et al*., 2008; Connaway and Dickey, 2011; VIAF, 2016): Proposed methods to create specific authority files for affiliation, publication venue, publisher and author names, respectively. They focused on data clustering methods, not in collecting data for enrichment.	We focused on collecting and extracting data for enriching a publication venue authority file.
Web collecting	(Pereira *et al*., 2008): Submitted queries to a web engine. However, they did not search need to reach specific pages, they collected data from the resulting snippets. (Pereira *et al*., 2009): Submitted queries to a web search engine to reach specific author pages. (Assis *et al*., 2008): Collected web pages using the focus crawler strategy.	Similar to Pereira *et al*. (2009), we submitted queries to a web search engine to reach specific pages, however the focus is on official publication venue pages for specific annual edition. Therefore, the strategies for document classification and data extraction are different. We did not use focus crawler as did Assis *et al*. (2008).
Classification	(Assis *et al*., 2008): Classified the collected web pages based on a genre and content approach.	Similarly, we evaluated a genre and content based classifier and differently we also evaluated a content only based classifier.
Data and information extraction	(Gunasundari and Karthikeyan, 2012; Abburu and Babu, 2013; Alfred *et al*., 2014): Proposed generic approaches for information extraction.	We used a simple and specific solution for extracting conference data through regular expressions and generated information from the sets of extracted data.

In previous works, Pereira *et al*. (2008; 2009) adopted the web information extraction strategy, through submission of queries to a search engine, to create a publication venue authority file and disambiguate author names, respectively. The strategy of using a search engine was also used similarly in this study. Other studies also explored data gathering using web search engines (Elmacioglu *et al*., 2007; Kan and Tan, 2008; Silva *et al*., 2009).

The approach presented by Assis *et al*. (2008) proposes a focused crawler that explores not only information related to the content but also genre, present on web pages. Thus, the crawling process can be expressed by two sets of terms, the first describing genre aspects (structure) of the desired pages and the second related to the subject or content. Our paper also investigates the classification based on genre and content to rank relevant pages of search engine results, since pages of publication venues usually have very similar structures, with terms of genre in common.

Information extraction from web pages is also a complex task. Several studies present proposals for it (Gunasundari and Karthikeyan, 2012; Abburu and Babu, 2013; Alfred *et al*., 2014). We chose a specific solution for extracting conference data through regular expressions.

Summarizing, our work is related to the following tasks: Authority file enrichment, web collecting, classification and data and information extraction. Table 1 shows how these tasks in the main works in the literature relate to ours.

PVAF

One of the contributions of the work Pereira *et al*. (2014) was the creation of a Publication Venue Authority File (PVAF) to the Computer Science field. This file consists of a set of records each one representing a publication venue. It stores the variant forms of writing the current titles and acronyms, the formerly titles and acronyms and other data such as ISSN, publisher, language, subject and bibliometrics, such as Qualis Capes and Impact Factor.

PVAF was created by collecting data from digital libraries and institutions that organize quality rankings of publication venues. It was created specific wrappers for each data source. The data in these sources are unstructured and often are incomplete and incorrect. Thus, manual adjustments were also necessary.

In their work, Pereira *et al*. (2014) also developed a method to search PVAF. It is a supervised learning method that uses PVAF to train a classifier, whose generated model is a set of association rules (Agrawal and Srikant, 1994) to identify publication venues. The association rules are of the form $X \rightarrow pv_i$, where X is a set of tokens and pv_i is a publication venue (e.g., $\{JCDL\} \rightarrow pv_1$). In the predict phase, it generates sets of tokens from the string to be searched, matches them with the antecedents of the rules in the model generated in the training phase and using a voting schema decides the prediction.

Figure 2 illustrates the PVAF schema. Data collected from the Web feed the PVAF database. These data are used to train an associative classifier whose generated learning model is used to predict user queries.

PVAF contains many acronyms and titles, however it does not have a matching among them, saying which acronym is associated with each title, both for current and formerly data. It also does not have information when there were changes in titles. The objective of this work is to get additional data from the Web to enrich the existent publication venue authority file.

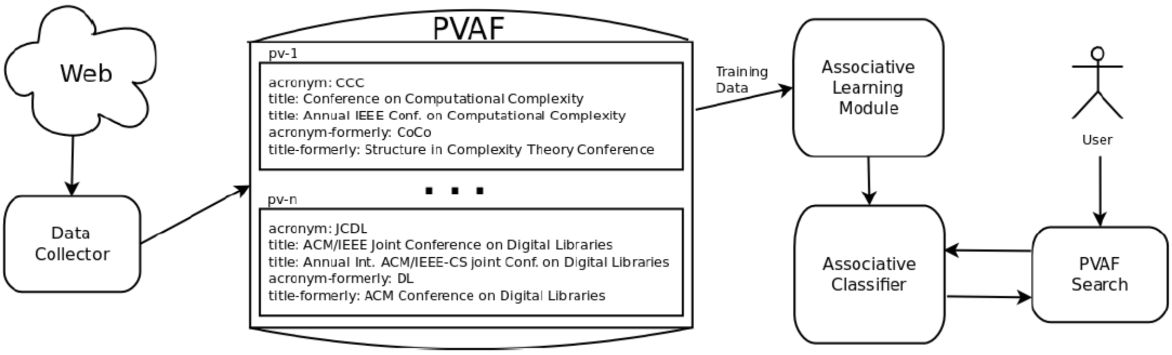

Web
Data Collector
PVAF
pv-1
acronym: CCC
title: Conference on Computational Complexity
title: Annual IEEE Conf. on Computational Complexity
acronym-formerly: CoCo
title-formerly: Structure in Complexity Theory Conference
pv-n
acronym: JCDL
title: ACM/IEEE Joint Conference on Digital Libraries
title: Annual Int. ACM/IEEE-CS Joint Conf. on Digital Libraries
acronym-formerly: DL
title-formerly: ACM Conference on Digital Libraries
Training Data
Associative Learning Module
Associative Classifier
User
PVAF Search

Fig. 2. PVAF schema

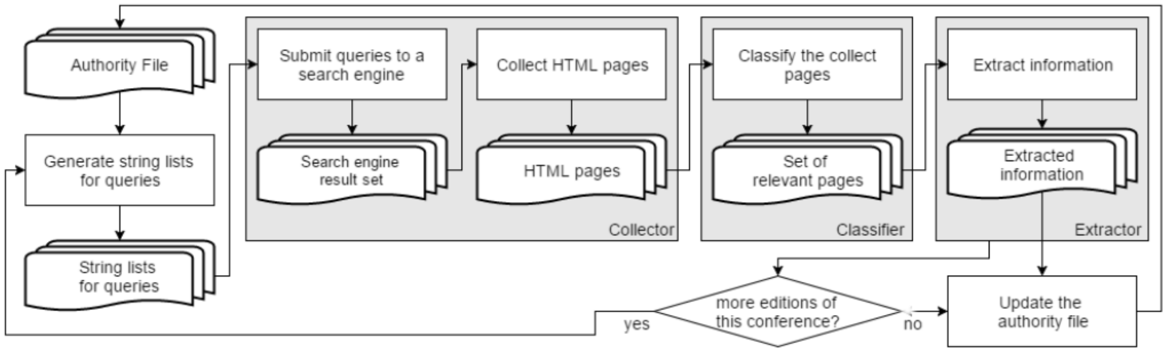


Fig. 3. Flowchart of our approach

Proposed Approach

The following is our approach to obtain information on conferences from the Web. Conferences are usually held in annual editions and each edition has its own web page. Figure 3 illustrates our approach.

The process starts by getting the basic data of conferences from the authority file. Using these data, strings are generated for queries in a search engine. Initially, queries are executed related to the current data of each conference using the current year. After that, the previous years are searched, reducing year by year until there is a period without finding relevant pages. A classifier is used to say whether a page is or is not relevant. Then, the system uses the data on formerly titles and acronyms of the conferences from PVAF in querying and these are now carried out at the first year in which no relevant page was found, reducing year by year again, until stay a certain period without finding any relevant page and so on for each formerly title and acronym.

Using the relevant pages collected, data extraction is made to complement the authority file. From the set of collected pages, it is also possible to trace a history of each conference edition with the years in which it occurred and when there was a change of name if it occurred.

For the best of our knowledge, works in the literature do not deal with automatic data collecting to feed or enrich an authority file. In the step of data collecting from the Web, we used the strategy to query a search engine, also used in other works (Pereira *et al*., 2008, 2009) and we adapted the queries in order to recover pages of conference editions. In the classification step, we evaluated the genre and content based classifier as in Assis *et al.* (2008) and we also evaluated a different approach based on content only. For data and information extraction, we implemented a specific strategy for publication venue, instead of using the generic strategies of other works (Gunasundari and Karthikeyan, 2012; Abburu and Babu, 2013; Alfred *et al*., 2014). In the following we present the details of each step of our approach.

Generating String Lists for Querying

Let $PV = \{pv_1, pv_2,..., pv_n\}$ be a set of n publication venues obtained from the authority file. For each $pv_i \in PV$, the sets $A = \{a_1, a_2, ..., a_o\}$, $T = \{t_1, t_2,..., t_p\}$, $AF = \{af_1, af_2, ..., af_q\}$ and $TF = \{tf_1, tf_2, ..., tf_r\}$ are obtained, where A, T, AF and TF are the sets of current acronyms and titles, formerly acronyms and titles of the publication venue, respectively.

For each $pv_i \in PV$, two string lists are generated, $L = \{l_1, l_2, ..., l_{o \cdot p}\}$ and $LF = \{lf_1, lf_2, ..., lf_{q \cdot r}\}$. List L is composed of strings with the combinations between sets *A* and *T* and list *LF* consists of strings with the combinations between sets *AF* and *TF*. This is necessary because there is no correspondence between acronyms and titles in the current authority file.

Querying a Web Search Engine

For each publication venue $pv_i \in PV$, queries to a search engine start with list *L*. At the end of each string in *L*, the algorithm adds the value of the current year and submits queries for each string in *L*. The query result sets are classified and when the classifier identifies the relevant page for that year, the algorithm decrements by one the year and submits new queries. If the classifier does not find a relevant page for a given year, it is marked and the search continues for previous years, up to a maximum of 5 years in a row. If in this period no relevant page is found, the algorithm returns to the marked year and starts to use the list *LF*, proceeding in the same way. After a period of five years without finding relevant pages, it concludes the queries for the current publication venue and starts the next pv_i.

The authority file may have oldest conference records which have been discontinued for more than five years. Then, it was established a period of 20 years to get the most recent edition. These values were obtained empirically.

For each item *i* in the list *L* or *LF*, the algorithm submits a query to the search engine and collects its top 10 results, forming a set $G_i = \{g_1, g_2,...,g_{10}\}$. Then the results of queries for each year are brought together in a single set *G*, which is sent to the classifier.

In our experiments, we used the Google search engine, which is recognized as the best web search engine and has shown good results in previous works (Pereira *et al*., 2008; 2009; Silva *et al*., 2009). Top 10 results were also used in these works.

Collecting and Preprocessing Web Pages

For each element $g_i \in G$ generated in the previous step, the collecting module accesses the URL of each page, collects and locally stores its HTML content. The collected pages are preprocessed for cleaning, which removes HTML tags, stop words, digits, special characters and converts uppercase characters to lowercase. It is generated a set of terms, where each word on the page is considered as a token that will be used by the classifier based on genre. Later the original HTML files will be used again for data extraction.

Classifying Relevant Pages

The classification step checks whether a page is relevant. A page is considered relevant when it is the official page of the conference for the searched year. For classifying pages, we used two approaches, one based on a combination of genre and content and the other based just on the content.

Classification based on genre checks the presence of the most common terms in a specific type of page, which usually define the page structure. In conference pages, for example, it is common terms of genre such as: Conference, submission, date, call, paper, keynote, speaker, committee, accommodation, program, among others. The terms of content, by the other side, are the terms that represent the specific publication venue, which is being considered for classification, for example, acronym, terms contained in the title and the conference year like in JCDL-ACM/IEEE Joint Conference on Digital Libraries - 2016.

The genre terms were generated in two ways: Manually by a human expert and automatically by a feature selection algorithm. The specialist generated its terms by observing a set of conference pages and listing their most common terms. The others were automatically generated for a feature selection algorithm, from that same set of pages.

The terms of genre were used as attributes for a feature vector in a Support Vector Machines (SVM) classifier (Vapnik, 1995), whose values were composed by the Term Frequency-Inverse Document Frequency (TF-IDF) (Baeza-Yates and Ribeiro-Neto, 2011) of the terms in the documents of the training collection. As a supervised classification technique, SVM requires training with documents classified by an expert. From this training, it generates a model that is used to classify other documents.

Classification based on content checks the page title, its URL and its content. For each publication venue, the algorithm generates a ranking of the pages by setting scores for each page collected. It looks for: (a) Year in the URL, (b) year in the page title, (c) year in the page content, (d) acronym in the URL, (e) acronym in the page title and (f) title in the page content. For the sake of efficiency of the algorithm, it considers only the first two thousand characters in the page content, which has been empirically verified that there is more likely to find items of interest. The algorithm also ensures that there is at least one mention to the title through exact term matching, or acronym of the publication venue and the year being searched. The algorithm counts the number of items found for each page and choose the one with the highest score. In case of tie, it chooses that one with the best position in the result set of the search engine, since it also has strategies to rank the most relevant pages in the top positions.

Classification based on genre and content obtains the intersection of the results of the classification based on genre and the classification based on content.

Extracting Data from Relevant Pages

An algorithm was developed to extract the following data from the relevant pages: URL, conference edition number, title officially used in that edition, acronym and date of occurrence.

For this, the algorithm uses regular expressions to search for patterns within HTML pages that allow the extraction of such data. How pages have similar formats, it is possible to obtain such data from many of them.

One of the most common patterns is formed by the edition number, followed by the publication venue title, acronym and year. For example, "41st International Conference on Very Large Data Bases (VLDB 2015)". For dates, the algorithm looks for patterns like "August 31 - September 4, 2015".

Extracting Information from the Set of Data

The modules presented in the previous sections were grouped in a framework capable of performing not only the data extraction from each web page, but also allows the information extraction from the sets of results obtained from multiple queries. If possible to obtain information on in which year a conference was held, when there was a change of name, which edition was held in which year and thus trace a history of the editions of each conference.

Experimental Evaluation

In this section we describe the datasets, metrics, experiments and their results.

Datasets

From PVAF, we obtained current and formerly acronyms and titles of conferences, which were used to generate two datasets, called C100 and CF107. These datasets were formed by metadata collected from search engine results, for data generated by the developed algorithm and the HTML files of the collected pages.

The C100 dataset consists of 100 conferences with edition in 2014, with queries regarding this year. The conferences were randomly selected from the authority file. For these conferences, we collected the results of the search engine and their HTML pages. All HTML pages were classified manually by a human expert. This dataset was created to evaluate the classifiers.

The CF107 dataset consists of 107 conferences that have in PVAF information about formerly acronyms and titles. For this dataset, the gathering was much larger than the previous one, because in this case we did not collect only one year for each conference, but several years from 2014, as the algorithm needed. The number of collected editions is specific to each conference. Using dataset CF107, it was possible to analyze and evaluate the results of the proposed approach and also identify the difficulties, problems and limitations of our approach.

Evaluation Metrics

In order to evaluate the classifiers we used the metrics precision, recall and F_1, given by:

$$Precision = \frac{|R \cap A|}{|A|}$$

$$Recall = \frac{|R \cap A|}{|R|}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where, $|A|$ is the number of pages classified as relevant by the classifier, $|R|$ is the number of pages classified as relevant by the specialist and $|R \cap A|$ is the number of pages classified as relevant by the classifier that are really relevant.

Experiments, Results and Discussions

Experiments with Dataset C100

The objective of the experiments with dataset C100 was to evaluate the proposed classification strategies. We evaluated several settings for the classifiers.

For classification based on genre, we used the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The experiments was performed using 10-folder cross-validation using the pages of the collected conferences. SVM parameters were adjusted by grid search, using a tool provided by LIBSVM.

The list of terms for classification based on genre was generated by a feature selection algorithm of Weka (Witten *et al*., 2011), as described following. The features were composed by tokens of all the pages manually classified as relevant and the feature values were the values of their TF-IDF. It was used the supervised attribute filter (weka.filters.supervised.attribute.AttributeSelection) to select attributes and the evaluator InfoGain (weka.attributeSelection.InfoGainAttributeEval), which evaluates the worth of an attribute by measuring the information gain.

The following 4 distinct settings for the classifier based on genre were evaluated:

- Configuration 1: For the page content, we considered all the terms contained therein and for the genre, we used the terms generated by a human expert, as shown in Table 2
- Configuration 2: For the page content, we considered all the terms contained therein and for the genre, we used the first 80 terms of the ranking generated by the feature selection algorithm, as shown in Table 2
- Configuration 3: For the page content, we considered only the content composed by the terms contained within the HTML tags <a> and </a>, which are commonly used in menus that make the link from the current page to other internal pages of the site. They are a very common structure used in conference pages and

the genre terms are usually located between these tags. In this configuration, we used the terms generated by the human expert

- Configuration 4: As in Configuration 3, but considering the terms generated by the feature selection algorithm

Table 3 shows the results obtained by the experiments. The classification based on content and the combination of genre and content are as described in Section Classifying Relevant Pages.

Comparing the results of the configurations for the classification based on genre, the values are statistically tied at 95% confidence level for the F_1 metric. This indicates that is not necessary a human expert to choose the terms of genre because the feature selection algorithm can get similar results automatically. In addition, it is not necessary to examine all page terms as those that appear in links also represent the genre content.

Similarly, it is observed that the results of the configurations for classification based on genre and content are also statistically tied for the F_1 metric. However, the results improved compared to the classification based on genre. This improvement is mainly due to higher values for the precision metric.

The classification based on content obtained such a good result as the classification based on genre and content for the F_1 metric. The difference is in the results of precision and recall.

Good precision is important to avoid adding incorrect data to enrich PVAF. Thus, we recommend using a classification based on genre and content for collecting data from a single year at a time. However, when we want to collect data from several years before the current, in order to obtain a history of the conference editions, the recall is important, as shown in experiments with the CF107 dataset that follows. In this case, the classification based on content is more suitable and is also more efficient.

Experiments with Dataset CF107

The objective of the experiments with dataset CF107 was to evaluate the data gathering of each edition of each conference in order to get their historical information, evaluating our approach as a whole, as we described in Section Proposed Approach, Fig. 3. This dataset consists of 107 conferences, which had name changing throughout their history, according to the PVAF data. We also evaluated the quality of data extraction from the relevant pages.

We performed two experiments. One to evaluate the best strategy for classification of relevant pages and the other to evaluate the effectiveness of our approach.

Table 2. Genre terms

Genre terms generated by a human expert
about, abstract, accepted, accommodation, author, award, call, chair, comittee, conference, date, deadline, edition, event, full, important, info, information, international, invited, keynote, local, notification, organization, organizer, paper, participant, partner, past, poster, presentation, proceeding, program, project, registration, research, scientific, scope, slide, speaker, sponsor, steering, student, submission, support, technical, tourism, travel, tutorial, university, venue, workshop
Genre terms generated by a feature selection algorithm
acceptance, accepted, accommodation, account, add, alert, algorithm, approach, article, association, author, book, browse, call, camera, chair, chapter, code, committee, conditions, conference, contact, content, date, deadline, demo, ebook, education, extended, feedback, held, help, home, hotel, important, instruction, invite, keynote, latest, material, message, notification, panel, paper, personal, policy, poster, pp, practitioner, print, privacy, professional, program, project, publication, ready, registration, researcher, resource, search, session, sign, speaker, sponsor, study, submission, submit, support, teaching, template, terms, together, travel, tutorial, venue, view, visa, vol, welcome, workshop

Table 3. Results of the classification experiments in dataset C100

Conf.	Precision (%)	Recall (%)	F_1 (%)
Classification based on genre			
1	85.9	78.6	82.1
2	84.6	78.6	81.4
3	89.8	74.0	81.1
4	87.9	77.5	82.4
Classification based on content			
-	88.8	87.0	87.9
Classification based on genre and content			
1	96.2	76.0	84.9
2	96.1	75.0	84.3
3	98.7	75.0	85.2
4	98.8	80.0	88.4

Table 4. Results of Experiment 1 for Classification Based on Genre (CBG), on Content (CBC) and on both (CBGC)

	Precision (%)	Recall (%)	F_1 (%)
CBG	94.2	53.1	67.9
CBC	90.2	68.7	78.0
CBGC	83.3	63.5	72.1

Table 5. Results of Experiment 2

	No. relevant pages	No. hits	Precision (%)
Current acronyms/titles	740	553	74.7
Formerly acronyms/titles	41	21	51.2
Total	781	574	73.5

Experiment 1

It was conducted on a random sample of six conferences in dataset CF107 using Configuration 4 above, which obtained the best results for classification in the experiments with dataset C100.

In total, data were collected for 153 possible editions of the conferences. Table 4 presents the results.

Good precision is important for a consistent data gathering, while a good recall allows the algorithm to retrieve more relevant pages and thus allowing that the previous editions of the conferences can be retrieved. Thus, the F_1 measurement shows that the classification based on content obtained the best results for the experiment.

Experiment 2

It was carried out on the full CF107 dataset, using the classification based on content, which achieved the best results in Experiment 1.

The algorithm analyzed 2,380 possible editions of the conferences in the dataset, totaling more than 32,400 pages collected. Table 5 presents the results. They are divided between pages obtained by querying involving the current and formerly acronyms and titles, according to the PVAF data. The table presents the number of pages classified as relevant by the algorithm, the number of pages correctly classified and the precision. The recall was not calculated due to the high number of pages to be checked manually.

Analyzing the details of the 207 cases of errors, we found that in 161 of them the search engine has not recovered the official conference page and the algorithm classified as relevant any other page that contained data on it. For the other 46 cases, the algorithm really missed the classification of the relevant pages.

The conferences with the highest number of editions identified by our approach generally have own domains (URLs) for their editions, with centralized management, making available the pages of old editions. By the other side, the conferences that had fewer editions identified have different domains for each edition. After some time such web pages become unavailable, remaining records of such editions only in digital libraries or third-party sites. In such cases, the approach actually did not identify the correct pages because they were unavailable or incorrectly identified other pages with very similar characteristics to conference pages.

Making a manual check in a sample of the conferences, we identified that in queries involving the formerly titles of conferences, there are some cases in which the pages exist, but the search engine did not recover them and in other cases, the pages really were not found through several different queries and manual searches.

Regarding to the data extraction, from the pages classified as relevant, the algorithm was able to extract the edition number from 142 of them, which appeared before the title of the conference. In this same position, it identified the year of occurrence from other 381 pages. The full date was extracted from 344 pages and the acronym from 17.

Regarding to the information extraction from the set of data, it was possible to identify some historical features on conferences. From the extraction of the year or edition number it was traced the history of each conference. For 74 of them, more than 5 editions were identified. Comparing the titles, it was possible to identify the year in which 22 of the conferences changed their names and 81 of them had an edition in 2014.

Difficulties, Problems and Limitations

The method has difficulty to classify pages that have very similar information to conferences, but which are not official pages of them. Such pages are usually sites of dissemination of scientific and related events and they are not interesting because may contain errors in data replication.

In older pages, it is common that the information in the page headers and in their menus are images, which is a problem to our classification process that in this implementation works only with text.

A limitation was found in the use of Google search engine API, which limits the number of queries that can be performed free of charge and then extends the time to collect the information.

Another limitation occurs when a conference page has been removed from the server, which occurs mainly for older editions. In addition, sometimes the server hosting a page was off-line at the time of collecting.

Conclusion

The results of this study demonstrate that it is possible to enrich an authority file of scientific conferences with information extracted from the Web. The proposed approach consists of the steps for collecting, classifying and extracting information from official pages of conferences.

The step for collecting pages is done by submitting queries to a web search engine. The search result sets need to be classified, to identify the relevant pages. We evaluated two strategies for classification: One based on content and another based on genre and content. The classification based on content was more effective to collect historical information. A simple extractor, based on regular expressions, was able to extract data as conference edition number, year, title, acronym and date of occurrence. From the extracted data, it was possible to trace a history of the editions of each conference.

We also evaluated the problems encountered and the main one is that older editions of many of the searched conferences do not seem to have their pages available on the web anymore. However, the pages of the most recent editions are easily found by the search engine.

As future work, the approach proposed for conferences will be adapted to journals and workshops. The structure of journal pages is different from conferences. Typically, journals have a single page, within which are their editions. Workshops mostly occur along with conferences and their information is usually contained on the same page of the associated conference. In addition, the extractor will be improved and extended to collect additional data such as the location of the event and the list of accepted papers. Such information is useful for data disambiguation.

Acknowledgement and Funding Information

This work was partially supported by the FAPEMIG grant CEX-APQ-01834-14, CNPq grant 200828/2015-0 and an individual scholarship from UFLA.

Author's Contributions

Heider Jesus: Design, implementation and manuscript writing.

Denilson Pereira: Research advisor and supervisor, Conception, design and manuscript writing.

Ethics

The authors confirm that this manuscript has not been published elsewhere and that no ethical issues are involved.

References

Abburu, S. and G.S. Babu, 2013. A framework for web information extraction and analysis. Int. J. Comput. Technol., 7: 574-579.

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, Sept. 12-15, Morgan Kaufmann Publishers Inc., Santiago, Chile, pp: 487-499.

Alfred, R., G.K. Soon, C.K. On and P. Anthony, 2014. A robust framework for web information extraction and retrieval. Int. J. Mach. Learn. Comput., 4: 146-150. DOI: 10.7763/IJMLC.2014.V4.403

Assis, G.T., A.H.F. Laender, A.S. da Silva and M.A. Gonçalves, 2008. The impact of term selection in genre-aware focused crawling. Proceedings of the ACM Symposium on Applied Computing, Mar. 16-20, ACM, Fortaleza, Ceara, Brazil, pp: 1158-1163. DOI: 10.1145/1363686.1363953

Auld, L., 1982. Authority control: An eight-year review. Library Resources Tech. Serv., 26: 319-330.

Baeza-Yates, R. and B. Ribeiro-Neto, 2011. Modern Information Retrieval: The Concepts and Technology behind Search. 2nd Edn., Addison-Wesley Professional, New York, ISBN-10: 0321416910, pp: 913.

Connaway, L.S. and T.J. Dickey, 2011. Publisher names in bibliographic data. Library Resources Tech. Services J., 55: 182-194.

Elmacioglu, E., M.Y. Kan, D. Lee and Y. Zhang, 2007. Web based linkage. Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management, Nov. 9-9, ACM, Lisbon, Portugal, pp: 121-128. DOI: 10.1145/1316902.1316922

French, J.C., A.L. Powell and E. Schulman, 2000. Using clustering strategies for creating authority files. J. Am. Society Inform. Sci., 51: 774-786. DOI: 10.1002/(SICI)1097-4571(2000)51:8<774::AID-ASI90>3.0.CO;2-P

Gunasundari, R. and S. Karthikeyan, 2012. A new approach for web information extraction. Int. J. Comput. Technol. Applic., 3: 211-215.

Kan, M.Y. and Y.F. Tan, 2008. Record matching in digital library metadata. Commun. ACM, 51: 91-94. DOI: 10.1145/1314215.1314231

Lee, D., J. Kang, P. Mitra, C.L. Giles and B.W. On, 2007. Are your citations clean? Commun. ACM, 50: 33-38. DOI: 10.1145/1323688.1323690

Pereira, D.A., E.E.B. da Silva and A.A.A. Esmin, 2014. Disambiguating publication venue titles using association rules. Proceedings of the IEEE/ACM Joint Conference on Digital Libraries, Sept. 8-12, IEEE Xplore Press, London, UK pp: 77-86. DOI: 10.1109/JCDL.2014.6970153

Pereira, D.A., B. Ribeiro-Neto, N. Ziviani and A.H.F. Laender, 2008. Using web information for creating publication venue authority files. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Jun. 16-20, ACM New York, NY, USA, pp: 295-304. DOI: 10.1145/1378889.1378940

Pereira, D.A., B. Ribeiro-Neto, N. Ziviani, A.H.F. Laender and M.A. Gonçalves *et al*., 2009. Using web information for author name disambiguation. Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, Jun. 15-19, ACM, Austin, USA., pp: 49-58. DOI: 10.1145/1555400.1555409

Silva, A.J., M.A. Gonçalves, A.H. Laender, M.A. Modesto and M. Cristo *et al*., 2009. Finding what is missing from a digital library: A case study in the computer science field. Inform. Process. Manage., 45: 380-391. DOI: 10.1016/j.ipm.2008.12.006

Sun, C.C., D.R., Shen, Y. Kou, T.Z. Nie and G. Yu, 2016. Topological features based entity disambiguation. J. Comput. Sci. Technol., 31: 1053-1068. DOI: 10.1007/s11390-016-1679-6

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. 2nd Edn., Springer, New York, ISBN-10: 0387945598, pp: 188.

VIAF, 2016. VIAF: The virtual international authority file.

Witten, I.H., E. Frank and M.A. Hall, 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann, Burlington, ISBN-10: 0080890369, pp: 664.