ABSTRACT

The automatic extraction of bibliographic data is still a difficult task to this day, when it is realized that the scientific publications are not in a standard format and each publications has its own template. There are many “regular expression” techniques and “supervised machine learning” techniques for extracting the complete details of the references mentioned in the bibliographic section. But there is no much difference in the percentage of their success.

This paper presents a strategy for segregating and automatically extracting the individual components of references such as Authors, Title of the references, publications details etc., using “Unsupervised technique” and link these references to their corresponding full text article with the help of google.

Keywords: Regular Expression, supervised machine learning, Bibliography, References, Unsupervised technique.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. Introduction

Researchers typically download and collect numerous research papers in PDF form onto their desktops for reading and for further reference. These downloaded PDF files are usually stored along with other files in our local file system.Reference manager softwares like RefWorks, Zotero, EndNote, SodhanaRef and Mendeley are available in the market take the help of extracted basic metadata such as Title, Author, Abstract, etc. to search a article. But the essential feature of “reference linking" is not focused as they are individual with no association or links even if a scholarly publication cites an already existing one in the researcher's personal computer, they are in no way associated and the researcher may not know the type of association between them.

Most of the researchers focused on document clustering based on their context. But this does not help a researcher who is actually performing literature survey. Classification helps in binding all the research articles that worked on the same area and suggest to the reader. Most of the researchers prefer the use of snowball sampling technique. In this particular technique, the researcher checks for the most relevant references in a primary article and tries to get the corresponding full texts of those references. Basically the first search will be on their personal computer. But to find a full text article in their file system is very time consuming and hectic task for the researcher. Instead they prefer to download from the web using the reference citation which causes redundancy. To make their search easy, there are many reference management software available in the market like Mendeley, Zotero, etc. which are good at extraction the metadata from each journal article and provide the researcher to find an article using its Title or Author name, etc. But still taking the title part of each and every reference and searching it manually in the reference management software is also a difficult task for researcher. To eliminate this problem and to reduce the manual strain for the researcher we have “SodhanaRef” which helps in automation of reference article linkage along with providing a semantic search for which the performance is 67% and also we have “A Strategy for

Automatically Extracting References from PDF Documents" in which used supervised technique with annotations and regular expressions are used. The increase in performance is 74% which is very less.So to improve the performance, we are demonstrating an unsupervised approach for automatically extracting the components of the references and thereby linking the references either to their file system(if the file is already present in their personal computer)or to the google scholar which helps in reducing the search time and manual strain for the researcher.

1.2. Machine Learning

Machine Learning is the ability to improve the behavior based on experiments and to learn from data with respect to some class of tasks and performance measures by choosing training data and how to represent the target function by choosing an algorithm to infer a target function.

The concept of Machine Leaning is categorized into :

1.Supervised Machine Learning

2.Unsupervised Machine Learning

1.2.1. Supervised Machine Learning:

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.

Supervised learning classified into two categories of algorithms:

Classification: A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".

Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

1.2.2. Unsupervised Machine Learning:

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information

without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore machine is restricted to find the hidden structure in unlabeled data by our-self.

Unsupervised learning classified into two categories of algorithms:

Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

1.2.3. Natural Language Processing:

Natural language processing (NLP) is a branch of Artificial Intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding. Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment anddetermine which parts are important.

While supervised and unsupervised learning, and specifically deep learning, are now widely used for modeling human language, there's also a need for syntactic and semantic understanding and domain expertise that are not necessarily present in these machine learning approaches. NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

1.3. Tools and Softwares

1. Natural Language Tool Kit
2. Python Programming Language

1.5. Motivation for Work

Researchers typically download and collect numerous research papers in PDF form onto their desktops for reading and for further reference. These downloaded PDF files are usually stored along with other files in the file system. These files are generally searched or referenced by their file names and it certainly depends on the user's memory recollecting power. By having the metadata of each file attached to it, the required file can be searched or retrieved by using this metadata. But this metadata was in no way attached to the downloaded file and never useful for further organization of files . So, the PDFs get lost in the depth of file system, as they are individuals with no associations and links. There are many methods that actually help the researchers to know the association between the research paper and database.

In proceedings, neither authors use nor editors check to guarantee that the adopted bibliographic templates were strictly followed. Problems often arise in the items in the list of references are incompleteness, or misclassification of existence of article. To avoid this we have "SodhanaRef" which helps in automation of reference article linkage along with providing semantic search for which the performance is 67%. So we are attempting to improve the performance using unsupervised approach for automatically extracting the components of references and thereby linking the references either to their file system or to google server.

1.6. Problem Statement

The task to this day, when it is realised that the scientific publications are not in a standard format and each publications has its own template. There are many "regular expression" techniques and "supervised machine learning techniques" for extracting the complete details of the references mentioned in the bibliographic section. But there is no much difference in the percentage of their success.This paper presents a strategy for segregating and extracting the individual components of references such as Authors, Title of the references, publications details etc., Using "Unsupervised technique" and link these references to their corresponding full text article

2. LITERATURE SURVEY

2.1. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-word corpus

Authors: P.D. Turney, M.L. Littman.

This paper has presented a general strategy for learning semantic orientation from semantic association. It could be argued that this is a supervised learning algorithm with some labeled training examples and millions or billions of unlabeled training examples. PMI-IR uses Pointwise Mutual Information and Information Retrieval to calculate the strength of the semantic association between words. Word co-occurrence statistics are obtained using IR. PMI-IR has been empirically evaluated using 80 synonym test questions from the Test of English as a Foreign Language (TOEFL), . For comparison, Latent Semantic Analysis (LSA) is used on the same 80 TOEFL questions. PMI by issuing queries to a search engine and noting the number of hits for this reason it is using AltaVista was chosen because it has a NEAR operator. SO-LSA applies Latent Semantic Analysis (LSA) to calculate the strength of the semantic association between words . LSA uses the Singular Value Decomposition (SVD) to analyze the statistical relationships among words in a corpus. Hatzivassiloglou and McKeown's (1997) set of 1,336 manually-labeled adjectives was not available for testing SO-PMI-IR and SO-LSA.

HM is restricted to adjectives, it requires labeled training data, and it is complex.HM is for automatically identify antonyms and distinguish near synonyms. Both synonyms and antonyms typically have strong semantic associations .HM has an accuracy of 78% and SO-PMI-IR has an accuracy of 80%. SO-PMI-IR requires a large corpus, but it is simple, easy to implement, unsupervised, and it is not restricted to adjectives.

2.2. An Approach Towards Establishing Reference Linking in Desktop Reference Manager

Authors: Mandava Kranthi Kiran, K. Thammi Reddy.

This paper presents for Extraction of authors and Separate titles from references list from journal article.Semantic search and reference linking depends

on the correctness of extraction metadata and the extraction of titles from the extracted references from the publication to provide a link to the already existing titles in XML/RDF store. The reference–article linkage is implemented with the help of semantic links concept to ease the process of snowball sampling, the title part of every reference in each and every article is extracted and automatically searched and linked to that particular full-text article whose metadata (which includes Title of the article, Author, Author details and References) has already been stored in XML/ RDF ¯le storage. If the journal article has a link then that corresponding full-text article is in his journal article collection. Reference-linking feature present for online navigation which helps in crawling from the citing article to the cited article be tracked from the article citing it, from a large volume of journal article collection .Snowball sampling technique is often used by a researcher while performing literature survey. But the difficulty arises when both the citing paper A and cited paper B are present in the standalone reference management software on a personal computer without the knowledge that A is citing B and B is downloaded again from the online digital library and stored in the reference management software causing redundancy and the effort that is spent in removing the duplicates. So, automation of detecting the cited article B that might be present locally and linking it to the citing article A is necessary. We need to extract the metadata from the PDF Journal, As PDF will have images, encrypted text, etc., we need a separate strategy of retrieving data from the PDF. So, we used ITEXT Library to help us in this task of extraction.

2.3. A Strategy for Automatically Extracting References from PDF

Authors: Neide Ferreira Alves ,Rafael Dueire Lins ,Maria Lencastre.

This paper presents a strategy for extracting references of scientific documents in PDF format. The scheme proposed was validated in LiveMemory platform, developed to generate digital libraries of proceedings of technical events. LiveMemory platform for extracting the list of references of a PDF document, which was digitally generated. Regular expressions, together with classification and identification based on K-NN algorithm and the Naïve Bayes algorithm were used for this purpose. In the case of electronically generated documents of formats such as PDF, PS, HTML and XML, the task of reference spotting is much easier,