

Pu-Jen Cheng, Hsin-Chen Chiao, Yi-Cheng Pan and Lee-Feng Chien, "Annotating text segments in documents for search," *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 2005, pp. 317-320.

doi: 10.1109/WI.2005.32

Abstract: It has been shown that annotating prominent text patterns contained in documents with appropriate types may benefit many applications. Most conventional tools for automatic text annotation extract named entities from texts and annotate them with information about persons, locations, dates and so on. However, this kind of entity type information is often short in length and is mostly limited to a small set of broader categories. In this paper, we try to remedy this problem by presenting an approach to extract global evidences from documents for improved named entity recognition. We also propose an unsupervised, generalized classification approach that collects training data from the Web automatically and classifies text patterns into more refined categories. Experimental results show the feasibility of the proposed approaches for search on the data of the NTCIR-2 information retrieval task.

keywords: {Internet;text analysis;NTCIR-2 information retrieval task;World Wide Web;automatic text annotation;named entity recognition;text pattern classification;text segment annotation;training data;unsupervised classification;Books;Data mining;Information management;Information retrieval;Information science;Infrared detectors;Noise robustness;Text categorization;Text recognition;Training data},

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1517864&isnumber=32502>

K. H. Yang and Y. H. Wu, "Author Name Disambiguation in Citations," *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, 2011, pp. 335-338.

doi: 10.1109/WI-IAT.2011.181

Abstract: Members of the academic community have increasingly turned to digital libraries to search for the latest work of their peers. On account of their role in the academic community, it is very important that these digital libraries collect citations in a consistent, accurate, and up-to-date manner, yet they do not correctly compile citations for myriads of authors for various reasons including authors with the same name, a problem known as the "name ambiguity problem." This problem occurs when multiple authors share the same name and particularly when names are simplified as in cases where names merely contain the first initial and the last name. In this paper, we propose an automatic algorithm to disambiguate authors in citations by using their specific attribute on the Web. A binary classifier and a cluster separator are applied to cluster citations. Our experiment results show that the disambiguation accuracy can be improved from 51% to 73% when using Web information.

keywords: {citation analysis;digital libraries;pattern classification;Web information;academic community;author name disambiguation;binary classifier;cluster citations;cluster separator;digital libraries;name ambiguity problem;Accuracy;Clustering algorithms;Libraries;Matched filters;Measurement;Noise;Particle separators;Author Name Disambiguation;Citation Grouping},

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6040873&isnumber=6036728>

A. Mansouri, L. S. Affendy and A. Mamat, "A New Fuzzy Support Vector Machine Method for Named Entity Recognition," *2008 International Conference on Computer Science and Information Technology*, Singapore, 2008, pp. 24-28.

doi: 10.1109/ICCSIT.2008.187

Abstract: Recognizing and extracting exact name entities, like Persons, Locations, Organizations, Dates and Times are very useful to mining information from electronics resources and text. Learning to extract these types of data is called Named Entity Recognition (NER) task. Proper named entity recognition and extraction is important to solve most problems in hot research area such as Question Answering and Summarization Systems, Information Retrieval and Information Extraction, Machine Translation, Video Annotation, Semantic Web Search and Bioinformatics. In this paper we have improved the precision in NER from text using the new proposed method that calls FSVM. In our method we have employed Support Vector Machine as one of the best machine learning algorithm for classification and contribute a new fuzzy membership function thus removing the Support Vector Machine's weakness points in NER precision and multi classification. The design of our method is a kind of One-Against-All multi classification technique to solve the traditional binary classifier in SVM.

keywords: {data mining;fuzzy set theory;information retrieval;pattern classification;support vector machines;text analysis;fuzzy membership function;fuzzy support vector machine;information mining;machine learning;named entity recognition;one-against-all multiclassification technique;Bioinformatics;Computer science;Data mining;Dictionaries;Information retrieval;Information technology;Semantic Web;Support vector machine classification;Support vector machines;Text recognition;Information Extraction;Information Retrieval;Named Entity Recognition and Extraction;Text retrieval},

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4624826&isnumber=4624813>