Intelligent Information Management, **2012, 4, 364-382**
http://dx.doi.org/10.4236/iim.2012.46041 Published Online November 2012 (http://www.SciRP.org/journal/iim)

Low-Power Themes Classifier (LPTC): A Human-Expert-Based Approach for Classification of Scientific Papers/Theses with Low-Power Theme

Mohsen Abasi[1], Mohammad Bagher Ghaznavi-Ghoushchi[2]
[1]Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran
[2]School of Engineering, Shahed University, Tehran, Iran
Email: abasi.mohsen@stu-mail.um.ac.ir, ghaznavi@shahed.ac.ir

Received September 1, 2012; revised October 26, 2012; accepted October 31, 2012

ABSTRACT

Document classification is widely applied in many scientific areas and academic environments, using NLP techniques and term extraction algorithms like CValue, TfIdf, TermEx, GlossEx, Weirdness and the others like. Nevertheless, they mainly have weaknesses in extracting most important terms when input text has not been rectified grammatically, or even has non-alphabetic methodical and math or chemical notations, and cross-domain inference of terms and phrases. In this paper, we propose a novel Text-Categorization and Term-Extraction method based on human-expert choice of classified categories. Papers are the training phase substances of the proposed algorithm. They have been already labeled with some scientific pre-defined field specific categories, by a human expert, especially one with high experiences and researches and surveys in the field. Our approach thereafter extracts (concept) terms of the labeled papers of each category and assigns all to the category. Categorization of test papers is then applied based on their extracted terms and further comparing with each category's terms. Besides, our approach will produce semantic enabled outputs that are useful for many goals such as knowledge bases and data sets complement of the Linked Data cloud and for semantic querying of them by some languages such as SparQL. Besides, further finding classified papers' gained topic or class will be easy by using URIs contained in the ontological outputs. The experimental results, comparing LPTC with five well-known term extraction algorithms by measuring precision and recall, show that categorization effectiveness can be achieved using our approach. In other words, the method LPTC is significantly superior to CValue, TfIdf, TermEx, GlossEx and Weirdness in the target study. As well, we conclude that higher number of papers for training, even higher precision we have.

Keywords: Natural Language Processing (NLP); Semantic Web; Term Extraction; Text Categorization; Resource Description Framework (RDF); Low-Power Theme

1. Introduction

There may be several survey papers in a research field. Latest and somehow old papers' highlights, in addition to a clear description of the field, are included in the survey papers. They summarize and organize recent research results and experiences in a novel way that integrates and adds understanding to tryings in the field. They use highlights of many papers in a research field to understand what sub-field each paper is in. Imagine an author want to write a survey paper or book in a research field. He has to read many papers have been accepted and indexed by many different conferences and journals. He must read them in details to understand accurately what research sub-filed each is in. Due to the large amount of papers, and also in many subfields, time and energy consuming is a typical process repeated each day until completion of the survey paper. So, automatic tools which can classify papers and help our imagined author in preparation of a good survey paper, are very useful and have an important value.

As another reason of using an automatic paper classifier, categorization of new papers submitted to a conference into the conference's research areas may be supposed. A conference chairman has to select a correct reviewer for each paper submitted to the conference in a little time. Since there are many large amounts of papers submitted to the conference, and they are in a lot of research fields covered in the conference, the process of selecting a correct reviewer has a lot of difficulties. In addition, selecting a correct reviewer has a vital importance in having a conference with a high impact factor; because a good paper may be rejected as a result of se-

