Utilising Wikipedia for Text Mining Applications

Muhammad Atif Qureshi
National University of Ireland, Galway
University of Milano-Bicocca
muhammad.qureshi@nuigalway.ie (matifq@gmail.com)
muhammad.qureshi@disco.unimib.it

7th October, 2015

Abstract

The process whereby inferences are made from textual data is broadly referred to as text mining. In order to ensure the quality and effectiveness of the derived inferences, several approaches have been proposed for different text mining applications. Among these applications, classifying a piece of text into pre-defined classes through the utilisation of training data falls into supervised approaches while arranging related documents or terms into clusters falls into unsupervised approaches. In both these approaches, processing is undertaken at the level of documents to make sense of text within those documents. Recent research efforts have begun exploring the role of knowledge bases in solving the various problems that arise in the domain of text mining. Of all the knowledge bases, Wikipedia on account of being one of the largest human-curated, online encyclopaedia has proven to be one of the most valuable resources in dealing with various problems in the domain of text mining. However, previous Wikipedia-based research efforts have not taken both Wikipedia categories and Wikipedia articles together as a source of information.

This thesis serves as a first step in eliminating this gap and throughout the contributions made in this thesis, we have shown the effectiveness of Wikipedia category-article structure for various text mining tasks. Wikipedia categories are organized in a taxonomical manner serving as semantic tags for Wikipedia articles and this provides a strong abstraction and expressive mode of knowledge representation. In this thesis, we explore the effectiveness of this mode of Wikipedia's expression (i.e., the category-article structure) via its application in the domains of text classification, subjectivity analysis (via a notion of "perspective" in news search), and keyword extraction.

First, we show the effectiveness of exploiting Wikipedia for two classification tasks i.e., 1- classifying the tweets[1] being relevant/irrelevant to an entity or brand, 2- classifying the tweets into different topical dimensions such as tweets related with workplace, innovation, etc. To do so, we define the notion of *relatedness* between the text in tweet and the information embedded within the Wikipedia category-article structure. Then, we present an application

[1]Message sent using Twitter.

in the area of news search by using the same notion of *relatedness* to show more information related to each search result highlighting the amount *perspective* or subjective bias in each returned result towards a certain opinion, topical drift, etc. Finally, we present a keyword extraction strategy using community detection over the Wikipedia categories to discover related keywords arranged in different communities.

The relationship between Wikipedia categories and articles is explored via a textual phrase matching framework whereby the starting point is textual phrases that match Wikipedia articles' titles/redirects. The Wikipedia articles for which a match occurs are then utilised by extraction of their associated categories, and these Wikipedia categories are used to derive various structural measures such as those relating to taxonomical depth and Wikipedia articles they contain. These measures are utilised in our proposed text classification, subjectivity analysis, and keyword extraction framework and the performance is analysed via extensive experimental evaluations. These experimental evaluations undertake comparisons with standard text mining approaches in the literature and our Wikipedia framework based on its category-article structure outperforms the standard text mining techniques.

Supervisor: Colm O'Riordan, National University of Ireland, Galway
Co-supervisor (minor role): Gabriella Pasi, University of Milano-Bicocca
Available: `https://aran.library.nuigalway.ie/xmlui/bitstream/handle/10379/5304/ARAN_opens_Access_to_Research_at_NUI_Galway.pdf?sequence=1`