Supervised Text Classification for News Filtering and Summarization on the Web

ABSTRACT
of
THESIS

SUBMITTED TO

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

LUCKNOW

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY

LUCKNOW

प्रज्ञा शील करुणा

ESTABLISHED 1996

FOR THE DEGREE OF

Doctor of Philosophy

IN

COMPUTER SCIENCE

Submitted by

Chandrakala Arya

Enrolment No. 957/13

Supervisor

Prof. Sanjay Kumar Dwivedi

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL FOR INFORMATION SCIENCE & TECHNOLOGY

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

(A CENTRAL UNIVERSITY; NAAC- 'A' GRADE)

VIDYA VIHAR, RAEBARELI ROAD, LUCKNOW-226 025 (U.P.), INDIA

2018

Abstract

With the rapid growth of broadcast systems, the internet and online information services, more and more information is available and accessible. Explosion of information has caused a well-recognized information overload problem. There is no time to read everything and yet we have to make critical decisions based on whatever information is available therefore the summarization of one or more documents become increasingly desirable. This problem was recognized and tackled early in 1950s. Since then several well-known algorithms have been developed, and until today it still remains an active research topic.

These days News Summarization has been identified as a crucial research area, Internet users are frequently turning to the web for news rather than going to traditional sources such as newspapers or television. Reading news online offers many benefits over traditional media. Thousands of news sources are available, and speed of access has improved so that even geographically remote sites are easily accessible.

News summarization is the process that allows readers more ability to control the quantity of text that they read. There are typically many articles on the same event. Summarization systems need to produce a concise and fluent summary conveying the key information in the input and help readers in determining if they want to access and read the full articles as well as allow them to get the idea of the reported event by reading the summary only. On the other hand Summarization is an ideal solution to provide condensed, informative document reorganization for faster and better representation of news evolution.

In news web page filtering and summarization, the task of news web page classification has remained in sharp focus since long. Classification is one of the supervised machines learning technique. Machine learning is a self-ruling system which is capable of acquiring and integrating knowledge constantly. For a classifier to learn how to classify the documents, it needs some kind of ground truth. For this purpose, the input objects are divided into training and testing data. A news web page classification phase classifies a news web page from a non-news web page. Prior knowledge of correct news page allows us to narrow our content extraction task considerably.

The motivation for this work comes from the need of retrieving useful web news pages from the Indian web news corpus. Web news pages differ from other web pages; it is especially important to identify web news accurately for correct classification. Our goal is to find a simple yet efficient method to extract news articles from web corpus. To do this, we propose an automatic recognition method that uses classification rules for web news based on a combination of content, structure and uniform resource locator (URL) attributes. News web articles have been gathered from 10 different news websites and used Naïve Bayes algorithm to distinguish news articles from non-news articles examples advertisements, not related links.

The experimental results demonstrate that the three classifiers built by Naïve Bayes, SMO and J48 all have a high precision for web news classification, while Naïve Bayes classifier shows high precision value for both set of experiments. Precision of Naïve Bayes classifier for the first set, when training and testing conducted on the dataset containing same websites using only two attributes, is 0.951 which is better than SMO and J48 as 0.859 and 0.766 respectively. On the other hand Naïve Bayes also shows the high precision value for the second set, when training and testing conducted on the dataset containing different websites for training and testing as 0.965, which is better than SMO and J48 as 0.847 and 0.817 respectively.

Web page filtering is used to let users to see only those portions of a page that are useful in summarization. News web page summarization face the two main issues: the first one is how to locate relevant documents (URL's) on the web and the second one is how to filter out irrelevant documents from a set of documents collected from the web. This work addresses the second issue. In our work for filtering stage we used the web information (mainly content) extraction, which retrieves the news webpages title and news content by using Tag Tree. In this work, we propose an approach for extracting the main content from news web pages. The proposed approach is based on the concept of tokenization of HTML page, these tokens construct the tag tree; web pages from different websites are parsed into Tag tree and generated a template from each web pages and discover matching patterns and multiple sequence alignment. It finds and removed shared token sequences from the web pages until the relevant information is extracted from them. We perform experiments on 500 web pages from ten different news websites. Experimental results show that the proposed approach efficiently extracts the relevant information and shows high value of precision, recall and F1-measure for

the all ten categories.

The task of news web pages filtering and summarization requires the extraction of important keyphrses from the news document. Keyphrase extraction from news web pages is an important task for news documents summarization. It has been a challenging research topic in recent years because news changes very rapidly. Only a small number of news websites have author given keyphrases and manually allocating keyphrases for each web news document is very effortful. Thus it is absolutely necessary to propose an approach for keyphrase extraction. Keyphrases are like index terms that enclose the important information about document content. Keyphrases actually offer concise and precise description of document content. Key phrases are considered as a single word or a combination of more than one word that represent the important concepts in a news article. The aim of this work is to develop and evaluate an automatic keyphrases extraction approach for news web pages. Our approach identifies the candidate keyphrases from documents and chooses those candidate keyphrase having highest weight score. Weight formula combines the feature set that includes TFIDF, phrase position in documents and lexical chain that is based on WordNet to represent semantic relations between words. The experimental results show that the performance of our approach is better than the contemporary approaches today.

The experiments indicate that the precision values of proposed approach when extracting 5 keyphrases are 0.34 which is 6.25% greater than KESR (0.32) and 21.4% greater than KEA (0.28). For extracting 10 keyphrase the precision of our approach is 0.22 which is 10% greater than KESR (0.20) and 15.7% greater than KEA (0.19) and finally for extracting 15 keyphrases the precision value of our approach (0.17) shows 5.6% lower value than KESR (0.18) while shows 13.3% higher results than KEA (0.15).

Recall value of proposed approach, KESR and KEA when extracting 5 keyphrases is 0.25, 0.24 and 0.29 respectively where, proposed approach shows higher improvement in recall values as 4.2% than KESR and 13.7% than KEA. When extracting 10 keyphrases the recall value of our approach is 0.46 which is 27.8% higher than KESR (0.36) and 15% higher than that of KEA (0.40). Recall value of all the three approaches when extracting 15 keyphrases are 0.51, 0.41 and 0.48 respectively whereas for the proposed approach shows 24.4% greater value than KESR and 6.25% greater value than KEA.

With the emergence of erroneous amount of online data, it is desirable to construct a news summarization approach that can extract, compare and rank sentences to create a summary of various news articles. In this work, we present a new method for news web page summarization based on similarity model and sentence ranking where most relevant sentences are extracted from the original news article. For sentence ranking, weight of the sentences has been computed by the combination of features direct keyphrase match, matching terms, sentence position, and sentence length. For the redundancy reduction, cosine similarity measure is used. We collect news articles from five different news websites representing the same event and topic written in English language. Experimental results show that our approach gives better results in comparison to other contemporary news summarizer approaches.

For testing the proposed approach, ten different categories of news articles namely Market, Business, India, Technology, National, Science & Environment, Politics, World, Entertainment, Sports has been collected from five famous Indian news websites (English) such as The Economic Times, The Hindu, The Times of India, Hindustan Times, and Indian Express. Each category has two sets and each set contains five news articles on the same topic related to the individual category. The proposed approach performs better than other baseline approaches TEES, SRRank, and LAKE.