Extraction of Weather Keywords from News Topic

Hnin Wut Yee, Myint Thu Zar Tun
Computer University, Maubin
nector128@gmail.com,myintthuzartun@gmail.com

Abstract

A large and growing number of web pages display contextual advertising based on keywords automatically extracted from the text of the page, and this is a substantial source of revenue supporting the web today. Despite the importance of this area, little formal, published research exists. We describe a system that learns how to extract keywords from web pages for advertisement targeting. The system uses a number of weather keywords presences in meta-data and how often the term occurs in search query. This system is trained with a set of example pages that have been hand-labeled with "relevant" keywords. Based on this training, it can extract new keywords from previously unseen pages.

Keywords: web pages, contextual advertising, substantial source, relevant keywords

1.Introduction

With more and more information flowing into our life, it is very important to lead people to gain more important information in time as short as possible. Keywords are a good solution, which give a brief summary of a web's content. With keywords, people can quickly find what they are most interested in and read them carefully. They will save us a lot of time. In addition, weather keywords are also useful to the research of the information retrieval, text clustering and topic search. Manually indexing keywords will cost highly. Thus, automatically indexing keywords from text is of great interests. [2]

News report is always a key way of information dissemination. The well developed of Internet causes that Web news becomes one of the most important channels which people acquire the newly-emerged things in the daily lives. However, great quantity of information on the Internet is reproduced, disseminated and stored. Consequently, the information on the Internet is highly susceptible to redundancy, noise, and inconsistent. Many document clustering and classification studies have been presented for browsing documents or organizing the retrieval results for easy viewing news articles.

News is always the main domain that people pay a large amount of attention to. Unfortunately, only a small fraction of documents in this field have keywords. However, compared to unrestricted text, news articles are relatively easy to extract keywords from, because they have the following characteristics. Firstly, a news document is always short in length, and usually only important words or phrases repeat. Secondly, as a rule, the purpose of news articles is to illustrate an event or a thing for readers. Then this kind of articles usually place more emphasis on some event entities such as weather conditions, places, organization and so on. Lastly, important content often occurs the first time in the title, or in the anterior part of the whole web page, especially the first paragraph or the first sentence in every paragraph. These characteristics will help us in keywords indexing.

Typical content-targeted weather systems analyze a web page such as a news page, weather page, organization page or another source of information to find weather keywords on that page. These keywords are then sent to a weather forecasting system which matches the keywords against a database of weather terms. Typically, if a user clicks on the weather terms, the forecaster is charged a fee, most of which is given to the web page owner, with a portion kept by the weather forecasting service.[1]

Picking appropriate keywords helps users in at least two ways: (i) choosing appropriate keywords can lead to users seeing weather keywords for knowledge of the weather terms, (ii)the better targeted the forecasting, the more revenue that is earned by the web page provider, and thus the more interesting the applications that can be supports. From the perspective of the forecaster, it is even more important to pick weather keywords.

The Topic Detection and Tracking (TDT) study intends to explore techniques for detecting the appearance of new keywords and for tracking the reappearance of them, and thus makes the people to handle the development of the weather news easily. The major task of the TDT study is the task of new segmentation: segmenting a continuous stream of broadcast weather news into distinct constituent weather news based on the events they describe.

Topic keyword extraction in weather news is the process of identifying the important keywords, in news that bear most of the topical content of weather news. For a long time, weather keywords extraction dominates the performance of weather news processing.

Moreover, some problems make the conventional document classification methods frequently with poor performance for weather news classifying. Web site news articles are always described depending on writing the habitual behavior of readers. Thus, the weather keywords are less apparent repeatedly, and furthermore the reporters may adapt different terms to descript the news event. In order to obtain high accuracy news classification results, to identify the weather keywords effectively become very important.

In this paper, we systematically investigated several different aspects of keyword extraction. We compared looking at each occurrence of a weather word or phrase in a document separately, versus combining all of our information about the word or phrase. Four Sections are illustrated as tag along. Section 2 illustrates the system architecture to express three steps. Section 3 explains the extraction process. In Section 4, we describe the experimental results and analysis of this system. Section 5 is given the conclusion of this paper.

2. System Architecture

During the extraction process , (i)a web page is taken as input, (ii) the web page is parsed and an attempt is made to detect the patterns and (iii) a number of weather new items (text/URL tuples) are produced as output.

Some simple data cleaning operations are performed on the news items in order to increase their quality with respect to data mining. News items can fit into more than one pattern During automatic extraction of news items, the initial strategy of our extraction tool is to detect items following the URL-text-URL pattern. This pattern has two identical links that function as news item delimiters, making the detection accurate. In this system, we use two types of features: (i) **Topic detection**: these properties give some evident notices for identifying the weather related keywords in weather news. (ii)**Characteristics of News Writing:** There are different kinds of writing In relation to news, there are only two major forms of writing: news writing and feature writing. News report portrays strictly on the occurrence and the course of a news event and so news writing is strictly based on news events. According to the definition of topic of TDT study, a topic is defined as "seminal activity or event, along with all directly related events and activities." In this process, there are two types of stages: Pre-processing stage and Topic Detection stage. The architecture design of the system is illustrated in Figure 1.

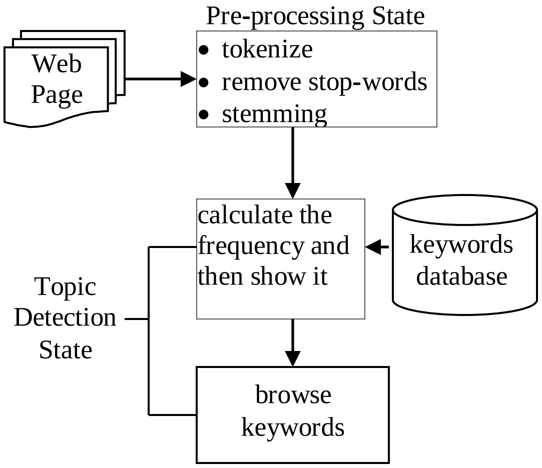


Figure 1. Architecture Design of the System

2.1 Preprocessing Stage

When a corpus (news) incomes into the system, this news is tokenized. *Tokenizing* means to separate words in this document. Coma and punctuation marks are not actual words. After tokenizing, stop-words are removed. *Stop-words* include articles, pronouns, some of the verbs, nouns and adjectives. For example, a, an, the, he, she, I, am, is, so, eat, and, very, to, etc. *Word stemming* means finding the root word of the given word, to make text processing more efficient. For example, replication, replicated, replica, replicate to replicate.

2.2 Topic Detection Stage

After pre-processing, there will be calculated by the topic detection method. It calculates the frequency of each candidate weather words and reserves the keyword words whose frequency. The keywords are extracted from the identified weather keywords which are the most discriminative weather keywords among the keywords in the news collection. According to the keywords, the weather keywords are classified for the performance of news processing.

3. Extraction Process

In this section, we introduce the general architecture of our keyword extraction system, which consists of the following three stages: recognizer module, filter module and selector module. This system is proved that better stop-word filtering of data can substantially improve the performance of the data processing. In the system of event tracking or news classification, keyword extraction strongly influences the accuracy of the classification. [1]

3.1 Recognizer Module

It can be seen that one document is composed of a set of character strings; every character string has its frequency in the document. In general, those character strings that occur several times can reflect the topic of the document. The main purpose of the recognizer module is to transform in HTML document into an easy-to-process plain-text based on document, while still maintaining important information.

In particular, we want to preserve the blocks in the original HTML document, but remove the HTML tags. For example, text in the same table should be placed together without tags like<table?, <tr>, or <td>. We also preserve information about which keywords are part of the anchor text of hypertext links. The meta section of an HTML document header is also an important source of useful information, even though most of the fields except the little are not displayed by web browsers. This module first parses an HTML document, and returns blocks of text in the body, hypertext information and meta information in the header.

3.2. Filter Module

So far, Weather character strings are generated by through frequency statistics. Thus, some of them stand out just because simple repetition and are probably not meaningful units of language.

The recognizer module treats equally all symbols in the text such as weather characters and punctuations, etc., Thus, when conduction the process of frequency statistics, for a character string, there might exist some punctuations and function words such as "storm". There punctuations and function words usually occur in the head or tail of a character string, It is evident that such character strings can't serve as keywords of an article and they should be deleted from the filter module.

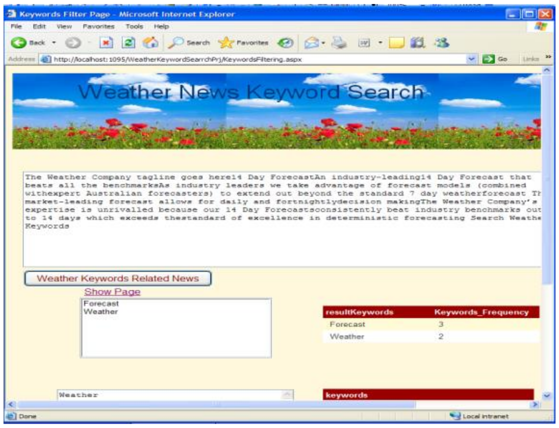
3.3 Selector Module

After filtering, now we can get a reduced set of filtering keywords. Most character strings in the set of meaningful and reflect the content of the document to some extent. For every word, now, we adopt several features to describe it.

We can find that the weather set is still too large to select from its keywords. Then we will conduct feature calculation to refine the candidate set. We have known that every processing item has a feature-value set. These feature values are our basis to evaluate every weather item. We can compute the score of every keyword. The higher the score, the more relevant the weather words are to the web page.

4. Experimental Results and Analysis

In this system, we select any web page in HTML file. During this selection, we choose view news button for web pages. After recognizer module, we parse this web page for modifying HTML code. The next step is filter module. This process removes the HTML source code and then tokenization process is followed to spite words. Then the system shows the number of weather keywords in the input web site as shown in Figure 2.


Keywords Filter Page - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://localhost:1095/WeatherKeywordSearchPrj/KeywordsFiltering.aspx
Go
Links
Weather News Keyword Search
Weather Keywords Related News
Show Page
Done
Local intranet

```
The Weather Company tagline goes here14 Day ForecastAn industry-leading14 Day Forecast that
beats all the benchmarksAs industry leaders we take advantage of forecast models (combined
withexpert Australian forecasters) to extend out beyond the standard 7 day weatherforecast Th
market-leading forecast allows for daily and fortnightlydecision makingThe Weather Company's
expertise is unrivalled because our 14 Day Forecastsconsistently beat industry benchmarks out
to 14 days which exceeds thestandard of excellence in deterministic forecasting Search Weathe
Keywords
```

Forecast
Weather

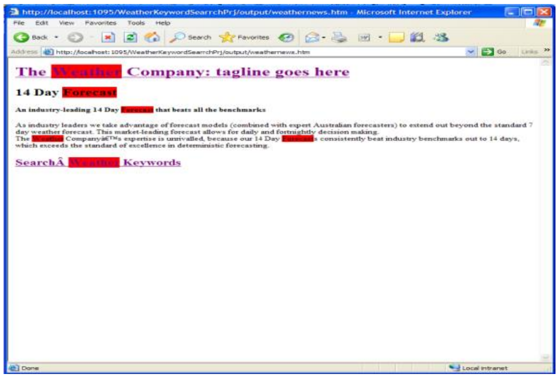
resultKeywords	Keywords_Frequency
Forecast	3
Weather	2

Weather

keywords

Figure 2. Weather Keywords Counts show in the System

Finally, there are weather keywords for information by using extraction process in selector module as shown in Figure3. Here, we automatically extracted keywords from them and evaluated the amount of keywords/ key phrases to show the user interface.



http://localhost:1095/WeatherKeywordSearrchPrj/output/weathernews.htm - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1095/WeatherKeywordSearrchPrj/output/weathernews.htm
Go Links
Done
Local intranet

The Weather Company: tagline goes here

14 Day Forecast

An industry-leading 14 Day Forecast that beats all the benchmarks

As industry leaders we take advantage of forecast models (combined with expert Australian forecasters) to extend out beyond the standard 7 day weather forecast. This market-leading forecast allows for daily and fortnightly decision making.
The Weather Companyâ€™s expertise is unrivalled, because our 14 Day Forecasts consistently beat industry benchmarks out to 14 days, which exceeds the standard of excellence in deterministic forecasting.

SearchÂ Weather Keywords

Figure 3. Weather Keywords show in Selector Module

5. Conclusion

Studies proved that better stop-word filtering of data can substantially improve the performance of the data processing. In the studies of event tracking or news classification, keyword extraction strongly influences the accuracy of the classification. Pre-processing stage can remove numbers meaningless terms from documents but have difficulty in obtaining the appropriate weather keywords to represent the weather news. This system considers and makes use of knowledge of the characteristics of news writing to develop a topic detection method for extracting the train data that are usually linked to the name-enteritis and place name typically to express actions in weather news.

References

[1]John O. C., Citing statements: "Computer recognition and use to improve retrieval", *Information Processing & Management.*, 18(3):125–131, 1982

[2]Yu-Sheng L., Chung-Hsien W., "Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology", ACM Transactions on Asian Language Information Processing (TALIP), Vol.1, No.1, March 2002, pp. 34-64.

[3]Turney, P.D., Learning to Extract Key-phrases from Text, NRC Technical Report ERB-1057, National Research Council, Canada, 1999.