Buletin Teknik Elektro dan Informatika
(Bulletin of Electrical Engineering and Informatics)
Vol. 2, No. 3, September 2013, pp. 177~181
ISSN: 2089-3191

Relevant Words Extraction Method in Text Mining

Naw Naw
Faculty of Informaton and Communication Technology, University of Technology
Yatanarpone Cyber City, Pyin Oo Lwin, Myanmar
e-mail: nawnaw1986@gmail.com

Abstract

Nowadays, E-commerce is very popular because of information explosion. Text mining is also important for information extraction. Users are more preferable to use the convenience system from many sources such as through web pages, email, social network and so on. This system proposed the relevant words extraction method for car recommendation system from user email. In relevant words extraction, this system proposed the Rule-based Technique based on Compiling Technique. Context- free grammar is very suitable for relevant words extraction. The extracted keys will be used in recommendation system. Recommendation System (RS) is a most popular tool that helps users to recommend according to their interests. In recommendation, this system proposed Content-based Filtering approach with Jaccard Coefficient that will help the users who want to buy the car by providing relevant car information.

Keyword: *relevant words extraction, rule-based, compiling technique, recommendation system, content-based filtering, Jaccard coefficient*

1. Introduction

Most E-commerce sites compelled to get the customers in various ways. Some used the information retrieval system. But this can only provide the information for user by matching with user's input. So this is not accurate. There are many constraints in information retrieval system. Users have the little choice in information retrieval system. So the ecommerce sites should let the users to input with free text as they like. The users do not need to worry about the structures and grammars. Most researches tried to solve this problem by text mining techniques. In Text mining, text is unstructured, amorphous, and contains information many different levels. Most text mining techniques are tried to extract useful data from unstructured text mostly written in natural language. The automatic extraction of information from text is to produce structured output that can be put into a database or others. Most systems use machine learning techniques and a variety of features such as Support Vector Machine, K means. Keywords are gathered, preprocessed and extracted based on corpus-oriented methods or document-oriented methods [3]. Some system use Rule-based technique. The accurate level between these systems is different. Rule–based technique gets 92% score. But Machine learning based approaches were able to achieve around 70% breakeven [9]. So this system proposed the rule-based text mining technique in extraction relevant words based on compiling technique. The output of the relevant words extraction method can be applied to the recommendation system. Recommender systems help customers to find what they really want. So this meets the requirements of customers in a short time. It helps users to find information, products, or by aggregating and analyzing suggestions from other users' activities. This proposed system intent to save time in extracting information from web application, to promote the performance of e-commence, to satisfy the customer dealing with car e-commerce, to help customers in finding product that they desired, to generate the new relevant words extraction method, to enhance the content based filtering method by combing with information extraction and to help businesses company for sale promotion.

2. Text Mining

Text mining uncovers the underlying themes or concepts that are contained in large document collections. Text mining applications have two phases: exploring the textual data for

Received January 6, 2013; Revised May 8, 2013; Accepted June 5, 2013

its content and then using discovered information to improve the existing processes. Both are important and can be referred to as descriptive mining and predictive mining. Descriptive mining involves discovering the themes and concepts that exist in a textual collection. Predictive modeling involves examining past data to predict future results [10]. In Descriptive mining, the unstructured texts are difficult to extract the useful data because of the richness and ambiguity of natural language. So this system proposed the relevant words extraction method based on context-free grammar. In predictive mining, this system proposed the content-based recommender system by using the output keys of relevant words extraction method.

3. Information Extraction

Text mining is powerful tool for the useful and valuable information extraction from huge data set [2]. An important approach to text mining involves the use of natural-language information extraction. Information extraction (IE) distills structured data or knowledge from un-structured text [4]. The proposed system will use the rule matching process from compiling technique. The context-free methods are powerful enough to describe almost all of the so-called syntactic features of programming languages. Indeed context-free grammars are often used in language manuals [5].

4. Recommendation System

The basic purpose of a RS is to recommend information items that will be of interest to a specific user. The most popular recommendation methods are [6]:

a. Content-based Filtering (CBF)
b. Collaborative Filtering (CF)
c. Rule-based Filtering (RBF)
d. Demographic Filtering (DF)
e. Case-based Reasoning (CBR)
f. Utility-based Filtering (UBF)
g. Knowledge-based Filtering (KBF)
h. Hybrid Approach (HA)

4.1. Content-based Filtering

This proposed system will use the content-based approach. CBF techniques are developed for information retrieval and information filtering research [7]. In the CBF system, each user can operate independently. In a content-based recommender systems, user will be recommended the most closely information of the items according to their request. In CBF, it is need to manipulate the similarity between the contents. There are many similarity methods used in content-based recommender system. But Jaccard Coefficient is most proper method for this proposed system.

5. Similarity Methods

There are many similarity methods to extract the information. They are cosine, Manhattan, Spearman's rank correlation coefficient, Kendall's rank correlation coefficients, Jaccard coefficient, Pearson and so on. Each of these has advantages and disadvantages respectively. This proposed system uses the Jaccard coefficient because it is suitable for the content based similarity.

5.1. Jaccard Coefficient

The Jaccard coefficient measures the similarity as the intersection divided by the union of the objects. The Jaccard coefficient is a similarity measure and ranges between 0 and 1. 1 means the two objects are the same and 0 means they are completely different. The nearer to 1 is, the more similarity between two items. Jaccard can be resolved these two conditions [8]:

a. The similarity should be maximal
b. There should be no similarity

$$S_{\text{Jaccard}} = \frac{MK}{TK} \quad (1)$$

Where,

MK= Match Keywords between sentence and database
TK= Total Keywords in a sentence

6. Related Works

Latha *et. al* [1] proposed Information Extraction from Biomedical Literature using Text Mining Framework. There are three steps in this paper. Text gathering: The documents are collected from the existing biomedical databases. Thousand-sample sets of documents are collected from various biological domains and these documents are analyzed and given as the input to the second stage. Text preprocessing stage: The above documents are preprocessed for decreasing the workload in the Data analysis stage. Data analysis: This phase focuses on analyzing the documents of the previous phase by using support vector machine (SVM). But this research wasted the time to recognized the every terms that are not concerned with biomedical information.

Ashwini Madane proposed Identifying Keywords and Key Phrases. A new algorithm (Kea) is used for automatically extracting key phrases from text. Step 1 (Preprocessing): stop word removing, tokenization. Step 2 (Candidate Identification): Kea then considers all the subsequences in each line and determines which of these suitable candidate phrases are. Step 3 (Determining Candidate Phrases): Use stemming method (Lovins). Step 4 (Feature Calculation): Kea builds a document frequency file. Use TF-IDF technique. But it takes too much time in candidate identification [2].

7. Implementation

Firstly, if the user sends the order from email, the system will extract the relevance words that are concern with their desired car information such as type of car, model number, amount of money they can afford, year, color, mileage, etc. We need the look up the main verb to divide the positive sentence or negative sentence. For example, I like the white color. In this sentence, this user wants the white color car. But there are negative sentences (e.g. I don't like the silver color). In this sentence, this user hates the silver color. So we need to distinguish the positive sentence and negative sentence by looking the main verb in a sentence. After the system gets the relevance keys from user email, content-based filtering approach will implement with Jaccard Coefficient method. This system proposed the Jaccard coefficient method with weight. The weight is very important in this system. Each user's emphasis is different. For example, some users may be emphatic the car model. But some users is more preferable the color than any others. So we need to find the user's most preferable things in the text to define the weight. Finally, the system will generate the recommendation list for that user according to their request.

7.1. Proposed Algorithm for Information Extraction

7.1.1. Sentence Extraction

Input : Email's content
Output : Relevance Sentences
Process :
Process for all Sentences
Sentence List ←Search the relevance sentence by comparing with candidate keys
End For

7.1.2. Key Extraction

Input : Sentences List
Output : Keys
Process :
Process Sentence-Level Identification
For each processed sentence

```
    Keys ← important-key-finder     (sentence, automobile-key)
  End For
Important-key-finder (sentence, automobile key)
  begin
    sub-sentence← stop word removal
     For each sub-sentence
      For each  RULE
        Rule matching process
          If matched rule then
            generate key-pair
          end If
       end For
     end For
  return generated key-pair
  end
```

When the system receives the user's email, the algorithm needs to decide which sentences are more concerned for the car recommendation. So we need to extract the most important sentences from the user's email. To know those sentences, we use the various keys that are related to the car model, color, price and so on. When we get these sentences, stop words removal process is performed. After stop words removing, we get the key phrases. These key phrases will be matched with proposed rules. Each rule proposed by the system can be added many more values. These values are dynamic.

7.1. Rules for Proposed System

Rule 1 ⇒ (< article >|< number >) < typeOf >

< article > ⇒ a | an | one | two | three | four | five | six | seven | eight | nine | ten

< number > ⇒ 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

< typeOf > ⇒ Toyota | Suzuki | Nissan | MarkII | car | Honda | Matsubishi | Isuzu | Subaru | Daihatsu | Mazda | Alfa | AlfaRomeo | Chrysler | Citroen | Fiat | Ford | GM | Hino | Mercedes Benz | Opel | Peugeot | Renault | Rover | Volkswagen | Volvo | BMW | Audi

Rule 2 ⇒ < prepositio n > < year >

< prepositio n > ⇒ for | in | at | since

< year > ⇒ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

⇒ < year > < year >

⇒ ε

Rule 3 ⇒< number >< notation >

< number >⇒ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

< number > ⇒ < number > < number >

⇒ ε

< notation >⇒ Lks | kyats | $

Rule 4 ⇒< color >

< color >⇒ pearl | white | black | grey | blue | light blue | dark blue | red | silver

Rule 5 ⇒< mileage >

< mileage >⇒ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

⇒ < milage >< mileage >

⇒ ε

Rule 6 ⇒< model number >

< model number >⇒ GRX120 | GX110 | JRX110 | CARINA ED | CAVALIER | CELICA | CENTURY | CHASER | COASTER | COROLLA | COROLLA 2 | UMAX_TOY | 86 | ALLEX | ALLION | ALPHARD | ALTEZZA | ALTEZZA WA GON | AQAU

8. Expected Outcomes

Today, our government opens the car market. So there are many demands on car. This system will help the user who wants to know the car information from their email. This system will recommend the closely relevant car information such as type of car, model, year, color, price and mileage for the requested user. Most information extraction systems use the machine learning technique. So they are very complex and time consuming. This proposed system can reduce these complexes by using Compiling technique. This system can decrease the preprocessing time with sentence level identification. Recommendation system performance will increase by combining this information extraction system.

References

[1] Latha K, Kalimuthu, Dr Rajaram R. Information Extraction from Biomedical Literature using Text Mining Framework. *IJISE*. USA. 2007.
[2] Ashwini Madane. Identifying Keywords and Key Phrases. IJSCE. 2012.
[3] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic Keyword Extraction from Individual Documents. 2010.
[4] Raymond J Mooney, Razvan Bunescu. Mining Knowledge from Text Using Information Extraction. 2009.
[5] PM Lewis II, DJ Rosenkrantz, RE Stearns. *Complier Design Theory*. ISBN 0-201-14455-7. 1978.
[6] Damian Fijalkowski, Radolsaw Zatoka. An Architecture of a Web Recommender System using Social Network User Profiles for E-commerce. IEEE. 2011: 287-290.
[7] Francesco Ricci. Content-Based Filtering and Hybrid Systems. 2005.
[8] Rares Vernica, Michael J Carey, Chan Li. Efficient Parallel Set-Similarity Joins Using MapReduce. 2010.
[9] Sundar Varadarajan, Kas Kasravi, Ronen Feldman. Text-Mining: Application Development Challenges. 2004.
[10] SAS Institute. Getting Started with SAS® Text Miner 4.1. ISBN 978-1-59994-999-4. 2009.
[11] Hany Mahgoub, Dietmar Rösner, Nabil Ismail, Fawzy Torkey. A Text Mining Technique Using Association Rules Extraction. *International Journal of Information and Mathematic Sciences.* 2008.
[12] Munyaradzi Chiwara, Mahmoud Al-Ayyoub, Mohammad Sajjad, Hossain, Rajan Gupta. CSE 634–Data Mining: Text Mining. 2009.
[13] Bernd Ludwig, Stefan Mandl. Centering Information Retrieval to the User. RSTI–RIA. 2010.
[14] BalaKrishna Kolluru, Sirintra Nakjang, Robert P Hirt, Anil Wipat, SophiaAnaniadou. Automatic extraction of microorganisms and their habitats fromfree text using text mining workflows. JIB, 2011.
[15] IAN H Witten. Adaptive Text Mining: Inferring Structure from Sequences. 2003.
[16] Yize Li, Jiazhong Nie, Yi Zhang, Bingqing Wang. Contextual Recommendation based on Text Mining. 2012.
[17] Haralampos Karanikas, Christos Tjortjis, Babis Theodoulidis. An Approach to Text Mining using Information Extraction. 2001.