Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 663 – 670

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

# Feature Extension for Short Text Categorization Using Frequent Term Sets

Yuan Man[a,b]*

[a]*China Huarong Asset Management CO., LTD., Beijing, China*
[b]*Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences , Univiersity of Chinese Academy of Sciences, Beijing, China*

**Abstract**

A short text feature extension method based on frequent term sets is proposed to overcome the drawbacks of the vector space model (VSM) on representing short text content. After defining the co-occurring and class orientation relations between terms, frequent term sets with identical class orientation are generated by calculating the support and confidence of word sets, and then taken as the background knowledge for short text feature extension. For each single term of the short text, the term sets containing this term are retrieved in the background knowledge and added into the original term vector as the feature extension. The experimental results on Sougou corpus show that the support and confidence have great impact on the scale of the background knowledge, but excessive extension also has redundancy and cannot obtain further improvement. The background knowledge based on frequent term sets is an effective way for feature extension. When the number of the training documents is limited, these extended features can greatly improve the classification results of SVM.



*Keywords:* frequent term sets, short text classification, feature extension

## 1. Main text

Text representation which refers to converting text content to a certain format for computer to process is a fundamental problem in text mining. The most popular text representation method in present is Vector Space

* Corresponding author. Tel.: +86-010-59618543; fax: +86-010-59618543.
*E-mail address:* yuanman@chamc.com.cn.


doi:10.1016/j.procs.2014.05.314

Model (VSM) [1]. In VSM, text contents are treated as Bag of Words (BOW), which ignores the associations of text features as well as the context and grammar structure. In recent years, following the emerging internet media such as social network and micro-blog, short text becomes an important type of text content online. Since the terms are rare in short text, it is weak for describing specific information, which aggravates the limits of VSM. To solve this problem, one of the method is to extend the text features to include more semantic, context and associations, using text mining technologies such as natural language processing, background knowledge and frequent item set mining.

Feature extension based on natural language processing is to use language models, grammar and syntax analysis to construct complex feature units [2-4]. Feature extension based on background knowledge tries to obtain more semantic information by retrieve the term in search engines [5,6], Wikipedia [7,8] or other outer resources. Frequent item set is a concept in association rule mining, and it also means frequent term set in text mining. It has been applied in text mining [9] because frequent term set reflect the associations of terms which involve more context and semantic information than single words. Cheng [10] analyzed frequent term sets for text categorization and proposed a frequent term sets selection method using multi-restraint metrics such as relevancy, coverage and redundancy. Ahone [11] firstly discussed the maximum frequent sequence (MFS) mining algorithm to avoid the duplication problem between each term set. Hernández [12] use MFS for text representation and each MFS corresponds to a text feature in the text vector space. A text clustering algorithm is then applied on this feature space.

This paper aims to improve short text representation in vector space model for short text categorization. A feature extension method based on background knowledge using frequent term sets is presented. Firstly, we define the co-occurring and class orientation relation of terms. Double term sets with co-occurring relation and identical class orientation relation are extracted from long text content to build the background knowledge. Then the original features are extended by the background knowledge and new features are added into the new feature space on which the SVM classifier is trained. Finally, experimental evaluation on real text data is conducted.

## 2. Feature extension based on background knowledge

Short text on Internet include multiple forms such as search key word, web review, micro-blog and news title. In this paper, we mainly focus on news title which appears frequently on social network feeds and shares, and the background knowledge are extracted from news content. The procedures involve: (1) text pre-processing, such as stemming and word segmentation; (2) mining double term sets to build the background knowledge; (3) feature extension using background knowledge and SVM classifier training; (4) feature extension on test data and evaluation of classification result.

### *2.1. Background knowledge based on frequent term sets*

Background knowledge are extracted from full content of document set $D=\{d_1, d_2, \ldots, d_n\}$ which is relevant to the short text. In document set $D$, term set $T=\{t_1, t_2, .., t_k\}$ is the collection of k terms, and $C=\{c_1, c_2, \ldots, c_m\}$ is the collection of class labels.

To select valuable frequent term sets, we consider the following restraints:

**Definition 1 (Support):** The support of term set $T$ is the number of documents which contain $T$ dividing the number of all the documents in data set , formulated as $sup(T) = Count(D_T) / Count(D)$. $D_T$ is the number of documents which contain $T$ and $Count(D_T)$ is the number of documents in $D_T$ and $Count(D)$ is the number of documents in $D$.

**Definition 2 (Confidence):** The confidence of association rule $t \Rightarrow c$ is *conf(t,c)*, which means the number of documents containing $t$ in class $c$ dividing the number of all documents involving t, formulated as