Expert Systems with Applications 38 (2011) 1393–1399

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Text stream clustering algorithm based on adaptive feature selection

Linghui Gong, Jianping Zeng *, Shiyong Zhang

School of Computer Science, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Keywords:
Text stream
Adaptive feature selection
Clustering

ABSTRACT

Text steam analysis is now of great importance and practical value today. It has several applications such as news group filtering, topic detection & tracking (TDT), user characterized recommendation etc. Clustering is one of the most important methods of analyzing text stream. However, most text stream clustering algorithms rarely consider the possible change of features during a long-time of clustering, which is usually the case, leading to unsatisfactory results of the clustering system. The paper mainly focuses on the problem of adaptive feature selection for clustering text stream. A validity index based method of adaptive feature selection is proposed, incorporating with which a new text stream clustering algorithm is developed. During the clustering process, threshold of cluster valid index is used to automatically trigger feature re-selection in order to ensure the validity of clustering. The experiment using Reuters-21578 text set as the text source shows that the clustering algorithm reaches reasonable results of high quality.

1. Introduction

With the fast popularization of Internet and great leap of network related technologies, Internet has changed people's lives worldwide, and Web 2.0 has changed the way we use Internet. Nowadays, people all round the world freely exchange information through Internet.

In the developing history of Internet, text information has been playing an extremely significant role. Today it is still the most fundamental and main form of information in the Internet. Therefore, the demand of supervising, managing text information and using it as valuable resource has increased a lot rapidly – text stream analysis is now of great importance and practical value.

Text stream analysis has several applications such as topic detection from a news stream, text crawling, document organization, topic detection & tracking (TDT), user characterized recommendation, user comments summary, trend analysis etc. The particularity of these analyzing tasks have in common is that text records come in the form of successive text sequence with time stamp. Result may be needed anytime as records everlastingly being generated.

Clustering is one of the most important methods of data mining (Han & Kamber, 2006). Actually, the clustering problem has recently been studied in the context of numeric data streams and categorical data streams (Aggarwal, Han, Wang, & Yu, 2003, 2004; O'Callaghan, Meyerson, Motwani, Mishra, & Guha, 2002).

Compared to traditional text clustering, in the text stream scene, challenges lie in several aspects: high algorithm efficiency is demanded in real-time; huge data set that cannot be kept in memory all at once; multiple scans from secondary storage is not desirable since it causes intolerable delays; and clustering algorithms need to be adaptive since data patterns change over time. Fig. 1 shows the difference between text stream clustering and the traditional one.

The main contributions of this paper are as follows. First, analyzing of feature selection algorithm employed in the traditional text clustering shows that static features are not suitable for the text stream context in the long-time condition. Second, a text stream clustering algorithm TSC-AFS (text stream clustering based on adaptive feature selection) is proposed based on adaptive feature selection strategy extended from the traditional algorithm. Third, a text stream clustering system using TSC-AFS is present and proves effective with experiment.

The organization of the paper is as follows. In the next section, related works are reviewed and the limitation of using unchanging feature set in text stream clustering. In Section 3, based on adaptive feature selection, we present a text stream clustering algorithm TSC-AFS. In Section 4, we evaluate the performance of TSC-AFS, in experiment and analyze the results. In the last section, we conclude the paper and point out the future research.

2. Related work

Feature selection is a process that chooses a subset from the original feature set according to some criterions, which is a sophisticated technology used in text mining. In text clustering, a text or

* Corresponding author.
E-mail addresses: daniel.gong.fudan@gmail.com (L. Gong), zeng_jian_ping@hotmail.com (J. Zeng), szhang@fudan.edu.cn (S. Zhang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.07.041

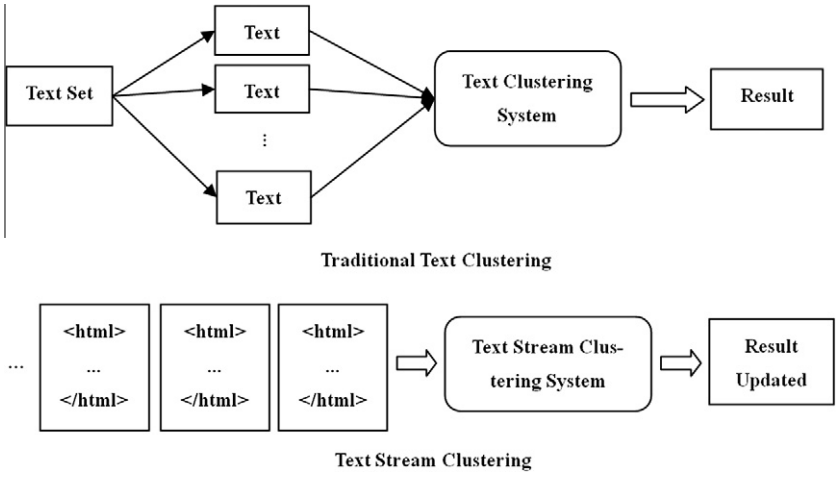

Fig. 1. Traditional text clustering vs. web text clustering.

document is usually represented as a bag of words, then dealing with the high dimension of the feature space results in huge computing complexity, which will surely lead to unsatisfying mining result. Feature selection, as an effective dimensionality reduction technique, reduces the bad impact to some degree, and produce better outcome.

As for feature selection for text clustering, there have some works on it. In the following part, we give a brief introduction on several feature selection methods. What's more, let D denote the document set, M the dimension of the features, and N the number of documents in the dataset.

2.1. Information gain (IG)

Information gain (Yang & Pederson, 1997) of a term measures the number of bits of information increasing for classification prediction because of the presence or absence of the term in a document or text. Let m be the number of clusters, c_i the ith cluster. The information gain of a term t is defined as

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum_{i=1}^{m} p(c_i|\bar{t})\log p(c_i|\bar{t}) \quad (1)$$

2.2. χ^2 statistic (CHI)

The χ^2 statistic measures the association between the term and the category (Galavotti, Sebastiani, & Simi, 2000). It is defined to be

$$\chi^2(t,c) = \frac{N\times(p(t,c)\times p(\bar{t},\bar{c}) - p(\bar{t},c)\times p(t,\bar{c}))^2}{p(t)\times p(\bar{t})\times p(c)\times p(\bar{c})}$$
$$\chi^2(t) = \underset{i=1}{\overset{m}{avg}}\{\chi^2(t,c_i)\} \quad (2)$$

2.3. Document frequency (DF)

Document frequency is the number of documents in which a term occurs in a dataset. It is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity. It is a simple but effective feature selection method for text categorization (Yang & Pederson, 1997).

2.4. Term strength

Term strength is originally proposed and evaluated for vocabulary reduction in text retrieval (Wilbur & Sirotkin, 1992), and later applied to text categorization (Yang, 1995). It is computed based on the conditional probability that a term occurs in the second half of a pair of related documents given that it occurs in the first half

$$TS(t) = p(t\in d_j | t\in d_i),\quad d_i,\quad d_j\in D\cap sim(d_i,d_j) > \beta \quad (3)$$

where β is the parameter to determine the related pairs. Since we need to calculate the similarity for each document pair, the time complexity of TS is quadratic to the number of documents. Because the class label information is not required, this method is also suitable for term reduction in text clustering.

2.5. Word variance-based selection

Word variance measures the ability of words to distinguish texts in the dataset. This algorithm sort all the words based on their variances and keep only the words with the highest variances (Zhong & Ghosh, 2005). That is, we set the size of feature set to be the same as the number of documents. The variance of the lth word is defined as

$$\sigma_l^2 = \frac{1}{N}\sum_x x^2(l) - \left(\frac{1}{N}\sum_x x(l)\right)^2 \quad (4)$$

where $x(l)$ is the number of occurrences of word w_l in document x.

As a basic method of data mining, clustering has been widely used in traditional text analysis because it requires almost no transcendental information of the object text set and produces practical category information. Fig. 2 shows the general construction of such a clustering system. However, the research on clustering of text data streams is still in early stage. Aggarwal and Yu presented an online algorithm framework based on traditional numeric data streams clustering with the use of a statistical summarization methodology for categorical and text data streams (Aggarwal & Yu, 2006). Yang raised a clustering method for online event detection of text data streams in a single pass (Yang, Pierce, & Carbonell, 1998). This method is also an extension of traditional numeric data streams clustering. Besides, Banerjee and Basu proposed several methods based on three batch topic models for clustering text data streams (Banerjee & Basu, 2007).

Actually, most existing clustering algorithms for text data streams are similarity-based approaches and often employ the