Data clustering of RSS content using hierarchical clustering method

Wen Hu, Qing he Pan

School of Computer and Information Engineering, Harbin University of Commerce, 150028
{Qing he Pan: 570749130@qq.com}

Abstract. Data clustering of RSS content is studied by using hierarchical clustering method. Based on a randomly chosen RSS list a words matrix is constructed. In this matrix each row represents a RSS title and stores the words that exist in the RSS according to a ratio. By using Pearson correlation coefficient the closeness between different RSSs content is computed and the result is used by hierarchical clustering algorithm. The tree-like graph is drawn to describe the hierarchical relationship.

Keywords: RSS, hierarchical clustering, Pearson correlation coefficient.

1 Introduction

RSS means “Really Simple Syndication”, “RDF (Resource Description Framework) Site Summary” or “Rich Site Summary”[1]. Actually all the three explanations indicate the same syndication technique. Now RSS is being wildly used in online news channels, blog and wiki. Using the RSS export of the website user can subscribe to the news and quickly obtain information.

Data clustering is a main method to finish this work. It a technique belongs to unsupervised learning techniques that are different from supervised learning techniques such as neural networks, decision trees and so on. The aim of data clustering is not to training with samples having right answers but to find out certain structures in object data.

In this paper we study the problem of RSS content clustering by using hierarchical clustering method. In section 2 we describe the format of the RSS data sets. In section 3 the hierarchical clustering is described and an algorithm is given. In section 4 an experiment is implemented using the algorithm. In section 5 we conclude this paper.

2 The data format

In order to execute the data clustering method the first step is to collect the RSS data and store them in certain way. In this research we collect data and store them in matrix. For a RSS url we first obtain all its content data. Then we strip all non-

alphanumeric characters, divide the data into separate word and count the times that each word occurring. So the matrix format is like table 1.

Table 1. The format of matrix segment

	word1	word2	word3	word4
RSS1	0	7	4	2
RSS2	9	3	7	6
RSS3	1	5	9	9
RSS4	5	2	2	8

It is easy to extract the words using regular expression. In order to deal with data easily all words are converted to lowercase. Based on this basic data format we can execute computation with appropriating method. In this paper we use hierarchical clustering method to analyzing data. With using this method the similar of different RSS is computed and compared.

We execute hierarchical clustering algorithm on data that has the format like table 1. It is obvious that algorithm analyzes the relationship between rows and it is our goal. But using a matrix transposition the algorithm can be applied to analyze and cluster the "words" without modification. In this paper we focus on the format in table1 and cluster different RSS content.

3 Hierarchical clustering

The hierarchical clustering method will always combine two most similar groups into a single group and construct a new hierarchical structure. In this research each group is composed of elements and each element is a RSS. The algorithm will compute the distances between every pair of groups and combine the two groups into a new group until there is only one group. This process can be depicted in figure 1.

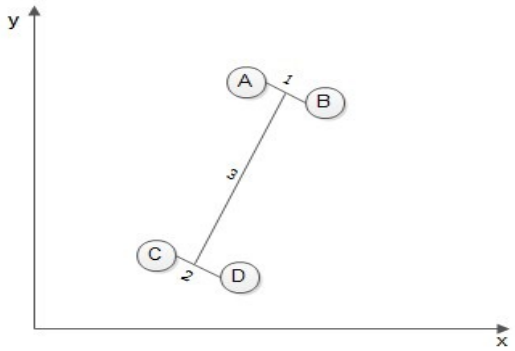

Fig. 1. The process of hierarchical clustering

Figure 1 depicts the process of hierarchical clustering. Suppose that the process starts from A. In the step1 we computes all distance between A and all other elements and the result indicates the distance between A and B is smallest, so A and B is

clustered. With the same way in step 2 C and D is clustered. Also in step 3 the groups formed by A B and C D are clustered together.

The distance that we use to evaluate the difference between groups can be computed by Euclidean distance, Pearson correlation coefficient, Manhattan distance, Kendall's (tau) distance or other method. In this research we choose Pearson correlation coefficient to evaluate the difference. The reason is that the different RSS has different number of words and Pearson correlation coefficient can correct this problem since it judges how well two different data sets fits onto a straight line[2]. The formula of Pearson correlation coefficient is given by (1)[3],

$$p(X,Y) = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (1)$$

In formula (1) X and Y represent two different vectors and in our research each corresponds to one row data in matrix. For example in matrix depicted in table1 X may described by vector (0, 7, 4, 2) that is the numbers of words contained in RSS1.

The hierarchical clustering algorithm can be designed by formula (1). We can use a vectors list to represent the matrix. Let n represent the vector number of M.

Algorithm hierarchical clustering(HC)

1 Input the data matrix M.
2 While n>1 :
2.1 Look for minimum of (P(m_i, m_j)), i !=j and m_i, $m_j \in M$. and set P(m_a, m_b) = min(P(m_i, m_j))
2.2 d_{min} = P(m_a, m_b)
2.3 Compute average value of m_a and m_b, avg(m_a, m_b)
2.4 Use avg (m_a, m_b) to form new cluster, new_cluster, m_a is its left child and m_b is its right child.
2.5 Delete m_a and m_b from M
2.6 Append new_cluster to M
2.7 Set n=len(M)
3 return M[0]

4 Experiment

In this section we do real data clustering using HC algorithm. The experiment has three main steps. The first step is to collect RSS data to construct the M matrix. The second step is to apply HC algorithm to M. The third step is to draw the hierarchical structure using the result in the second step. RSS data is collected by Universal Feed Parser[4]. In this research we gather 25 RSS. For each RSS we collect its data, use these data construct M matrix and save M as a .txt file. Apply HC to M and get the tree-like graph in figure 2.

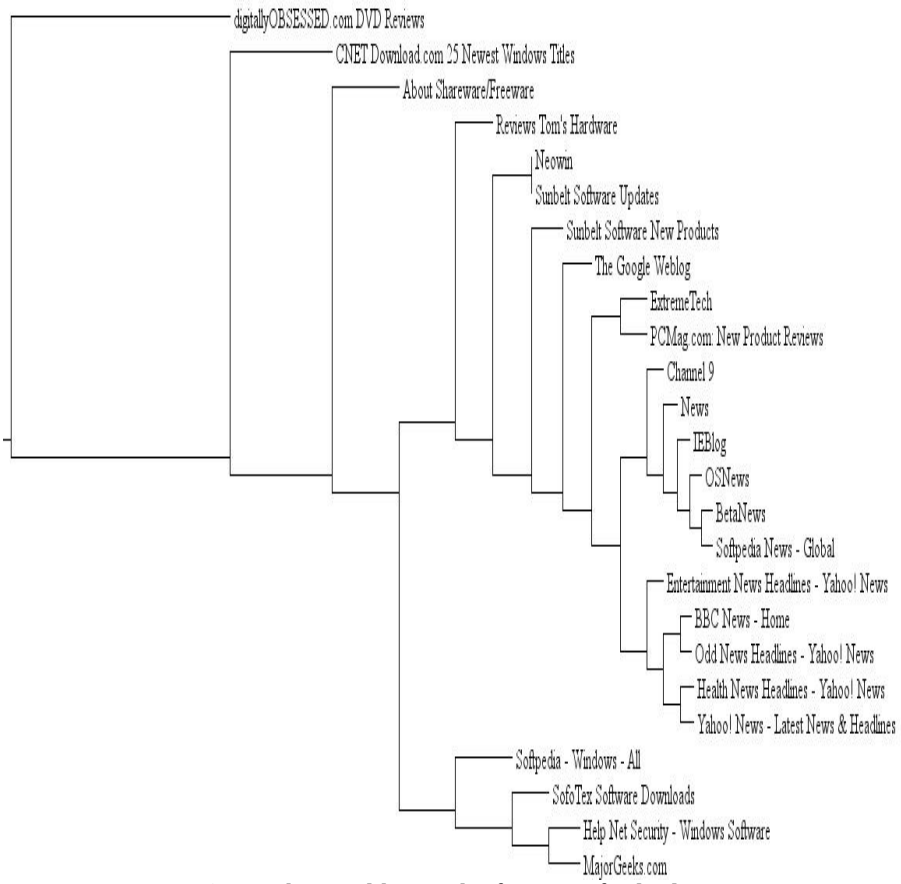

Fig. 2. The tree-like graph of 25 RSS feeds clustering

5 Conclusion

In this paper we illustrate how to cluster RSS data using hierarchical clustering method. Through clustering process analysis we describe the hierarchical clustering algorithm. By applying this algorithm to a real clustering problem we prove its effectiveness.

References

1. http://en.wikipedia.org/wiki/RSS
2. Toby Segaran. Programming Collective Intelligence. 2007, O'Reilly.
3. http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
4. http://www.feedparser.org