Clustering IP Addresses Using Longest Prefix Matching and Nearest Neighbor Algorithms

Asim Karim, Syed Imran Jami, Irfan Ahmad, Mansoor Sarwar, and Zartash Uzmi

Dept. of Computer Science
Lahore University of Management Sciences
Opposite Sector U, DHA, Lahore, 54792, Pakistan
akarim@lums.edu.pk

Abstract. This paper summarizes a new algorithm for clustering IP addresses. Unlike popular clustering algorithms such as k-means and DBSCAN, this algorithm is designed specifically for IP addresses. In particular, the algorithm employs the longest prefix match as a similarity metric and uses an adaptation of the nearest neighbor algorithm for search to yield meaningful clusters. The algorithm is automatic in that it does not require any input parameters. When applied to a large IP address dataset, the algorithm produced 90% correct clusters. Correct cluster analysis is essential for many network design and management tasks including design of web caches and server replications.

1 Background

Clustering is a key task in the discovery of useful patterns in large datasets. Clustering algorithms divide the data objects in the dataset into disjoint sets such that the objects within a set are more similar than to the objects in other sets. Over the years, many clustering algorithms have been developed employing various similarity metrics and search heuristics [1]. In general, these algorithms are general-purpose data clustering techniques that rely on domain-independent similarity metrics and search heuristics.

Internet protocol (IP) addresses are universally used for computer network communication today. The analysis of IP addresses contained within network traffic flows can yield useful patterns for traffic engineering such as the design of web caches and server replications. Clustering is an attractive technique for segmenting network traffic flows based on IP addresses. However, popular clustering algorithms such as k-means, k-medoids, and DBSCAN [1] do not produce meaningful clusters when applied to IP addresses [2].

2 Our Algorithm

We have developed a new algorithm for clustering large IP address datasets that uses the longest prefix match as the similarity metric and an adaptation of the nearest neighbor heuristic for clustering. This is a domain-specific algorithm that takes into consideration the unique characteristics of IP addresses. An IP address can be repre-

C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004, LNAI 3157, pp. 965–966, 2004.

sented by a 32-bit-long string. The longest prefix match between two IP addresses is the largest number of prefix bits that are identical in the two addresses [3]. This concept is used to determine similarity between IP addresses; the larger the longest prefix match the greater the similarity and likelihood that the addresses belong to the same network domain [2].

The nearest neighbor clustering algorithm merges a data object into the existing cluster to which it is the most similar provided the similarity is greater than a pre-specified threshold value; otherwise, it is created as a new cluster [1]. Our algorithm adapts the nearest neighbor algorithm by using the longest prefix match as the similarity metric and eliminating the need for a threshold value to be pre-specified.

The new algorithm for clustering IP addresses is summarized next. First, the longest prefix match among the IP addresses in the dataset is calculated and stored in an adjacency matrix. Then, each IP address is considered in turn and its cluster is created with all IP addresses with which it has the largest longest prefix match. In other words, the nearest neighbor concept is applied. However, unlike in the original nearest neighbor algorithm, a new cluster is created for every IP address with the IP addresses with which it has the largest longest prefix match. As such, IP addresses may be relocated from one cluster to another whenever their longest prefix match is greater with another IP address. In this way, clusters are modified iteratively as each IP address is considered based on the longest prefix match, a natural measure of similarity for IP addresses. Notice that our algorithm does not require the input of a threshold value for the similarity, as required in the original nearest neighbor algorithm. This makes the algorithm automatic.

3 Results

The algorithm is tested on a dataset containing 10,525 distinct IP addresses. The clustering results are verified by using domain name lookup (nslookup) utilities [4]. It is found that about 90% of the clusters formed by the algorithm are valid clusters representing natural groups of IP addresses. In other words, the algorithm is able to find clusters of IP addresses belonging to the same network domain in almost all cases.

References

1. Maragaret H. Dunham, “Data Mining: Introductory and Advanced Topics”, Pearson Education, 2003
2. Balachander Krishnamurthy, Jia Wang, “On Network-Aware Clustering of Web Clients”, ACM SIGCOMM ‘00, Stockholm, Sweden, 2000
3. Marcel Waldvogel, “Fast Longest Prefix Matching: Algorithms, Analysis, and Applications”, Swiss Federal Institute of Technology, Zurich, http://marcel.wanda.ch/Publications/waldvogel00fast.pdf
4. NS lookup Utility, http://ws.arin.net/cgi-bin/whois.pl