A Comparative Study of Clustering Algorithms

Kanna Al Falahi, United Arab Emirates University-Al Ain, UAE

Saad Harous, United Arab Emirates University-Al Ain, UAE

Yacine Atif, United Arab Emirates University-Al Ain, UAE

ABSTRACT

Clustering is a major problem when dealing with organizing and dividing data. There are multiple algorithms proposed to handle this issue in many scientific areas such as classifications, community detection and collaborative filtering. The need for clustering arises in Social Networks where huge data generated daily and different relations are established between users. The ability to find groups of interest in a network can help in many aspects to provide different services such as targeted advertisements. The authors surveyed different clustering algorithms from three different clustering groups: Hierarchical, Partitional, and Density-based algorithms. They then discuss and compare these algorithms from social web point view and show their strength and weaknesses in handling social web data. They also use a case study to support our finding by applying two clustering algorithms on articles collected from Delicious.com and discussing the different groups generated by each algorithm.

Keywords: *Algorithms, Clustering Algorithms, Community Detection, Data, Social Networks, Social Web Point View*

INTRODUCTION

The web is a huge repository of information of all kinds and types. Through the years, the web has evolved dramatically. From the static Web 1.0, where webmasters create and upload web pages with limited interaction possibilities, to the more dynamic Web 2.0, where contents are collaboratively generated and communicated across blogs, feeds and social networks. The advent of Web 3.0 brought more intelligence to Web contents through the evolution of the Semantic Web and more automation of services over the Web to further support machine-to-machine interactions (Wikipedia, 2010c).

The Semantic Web provides novel models for retrieving and analyzing Web information. Intelligent Web applications are emerging to analyze users' inputs, behaviors and respond accordingly to different contextual considerations. For example, what if you use a Web application to learn about Programming. The Web application would realize your experiential-learning style from your electronic profile and guides you along personalized instructional material that best meet your learning style. Semantic Web-based applications analyze interactions

DOI: 10.4018/jvcsn.2011070101

and profile users based on past history or pre-established records. Another possibility follows a case-based approach to match users with similar assets and aspirations to common Web experiences. The opportunity to analyze similarities within social context empowers Web experiences through identifying the commons to recommend preferential Web contents and services (Adomavicius & Tuzhilin, 2005).

Connectivity is a core feature of the above intelligent Web applications, where users share files, publish articles, comment on others' blogs or forums, view users' profiles and add new members to their connections. These are typical operations within today's social networks such as Facebook, MySpace and Twitter. To make useful inferences over social connections, intelligent Web applications need three typical knowledge-based modules (Marmanis & Babenko, 2009):

1. **Content**: represented by the hypermedia data of the considered domain and composed of inter-linked resources.
2. **Reference**: or the knowledge-base that tags and annotates domain content through rules, which categorize contents into meaningful folksonomies (Anfinnsen et al., 2010).
3. **Algorithms**: which run the inference engine modules on aggregated content.

People feed the Web with information every day. This continuous flow of information may result in some inconsistencies, as users will have myriad choices that need to be organized in an efficient manner. Data classification and clustering facilitate the process of analyzing and building meaningful inferences for example grouping similar Web pages could reveal serious problems such as mirrored Web pages or copyright violation (Haveliwala et al., 2000). In the intelligent Web, there are two algorithmic approaches to categorize data: Clustering and Classification (Marmanis & Babenko, 2009). These approaches are useful in performing targeted advertisements or personalizing Web experiences by allowing users to view posts that specifically interest them (Adomavicius & Tuzhilin, 2005) such as special content recommendation or page categorization (like Google News).

The objective of this paper is to provide means to identify individuals and data groups in the Web that are relevant to a given user. We focus on clustering algorithms in social Web context, particularly Hierarchical, Partitional and Density-based algorithms. We also discuss and compare six important algorithms used for this purpose namely: Link-based (Single-Link, Average-Link and MST Single-Link), K-means, ROCK and DBSCAN algorithms.

The rest of the paper is organized as follows: First, we define clustering and then, a section that discusses the different types of clustering techniques. We then introduce some terms and concepts related to clustering processes and provide an overview of different clustering algorithms. Following that we compare between the clustering algorithms presented in this paper. Next we illustrate a case study related to using clustering algorithms in social networks context and finally, we conclude the paper with a summary of work and suggestions for further future extensions.

CLUSTERING IN SOCIAL WEB

Creating associations among people in the form of groups is one of human natures. Previously, people used clustering in order to study phenomena and compare them to other phenomena based on a certain set of rules. Clustering refers to grouping similar things together. It is a classification of data or individuals into groups of similar instances. Each group is called a cluster. It consists of entities that embody some similarities and are dissimilar to entities belonging to other groups (Berkhin, 2002). We can find many definitions for clustering in the literature (Jain et al., 1999; Xu & Wunsch, 2005; Gower, 1971; Jain & Dubes, 1988; Mocian, 2009; Tan et al., 2005), but the most common definition is partitioning data or individuals sets into groups (called clusters), based on some crite-