CS999: Web Data and Text Mining
Annotated Bibliography

Justin Kamerman 3335272

February 9, 2011

Bibliography

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," *KNOWLEDGE AND INFORMATION SYSTEMS*, vol. 1, pp. 5–32, 1999.

[2] A. Ghorbani and I.-V. Onut, *Y-Means: An Autonomous Clustering Algorithm*, ser. Hybrid Artificial Intelligence Systems. Springer Berlin / Heidelberg, 2010, vol. 6076, pp. 1–13.

[3] C. H. and Z. D., "Ai and opinion mining," *IEEE Intell.Syst.IEEE Intelligent Systems*, vol. 25, no. 3, pp. 74–76, 2010.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput.Surv.*, vol. 31, no. 3, pp. 264–323, September 1999. [Online]. Available: http://doi.acm.org.proxy.hil.unb.ca/10.1145/331499.331504

Paper is an overview of data clustering concepts and techniques. Clustering is an exploratory undertaking (unsupervised), as opposed to classification (supervised). During clustering, a collection of patterns are organised based on a notion of their similarity to one another. Patterns are typically represented as feature vectors and their organisation occurs within this feature space. Pattern selection involves feature selection as well as feature extraction, transforming input features to produce new salient features. All clustering algorithms will produce clusters regardless of the underlying data so how do we evaluate a cluster algorithm ? Cluster validation studies can be *external*, comparing the recovered structure to an *a priori* structure; *internal*, which examines whether the structure is intrinsically appropriate for the data; or *relative*, which compares two structures and measures their relative merit.

A measure of the similarity between two patterns is essential to most clustering procedures. The most common measure is the Euclidean distance. It works well when a data set has compact, isolated clusters but large scale features tend to dominate unless weighted or normalized. Salient cluster algorithm properties include: agglomerative vs divisive; monothetic vs polythetic; hard vs fuzzy; deterministic vs stochastic; incremental vs non-incremental; and hierarchical vs partitioning.

Hierarchical algorithms produce a *dendrogram* representing nexted groupings of patterns and the similarity thresholds at which they change. Most hierarchical cluster algorithms are variants of the single-link, *single-link* (distance between clusters is the minimum between any two patterns drawn from different clusters); *complete-link* (distance between clusters is the maximum between any two patterns from different clusters); and *minimum-variance* algorithms. Partitioning algorithms are less demanding computationally compared to hierarchical algorithms. A problem of these algorithms is the choice of the number of desired output clusters. The most common and intuitive criterion function used in partitional clustering is the *squared-error* criterion of which *K-Means* is the simplest and most commony used. *K-Means* is one of the most efficient in terms of execution time and one of the few methods appropriate for use on large data sets. *K-Means* requires one to specify the number of clusters to create which is difficult to do optimally. Variants of *K-Means* have been proposed which dynamically merge and/or spilt clusters based on a variance threshold.

Clusters are typically represented by their centroid, a simple scheme if clusters are compact and iso-tropic. If clusters are elongated or non-isotropic, then this representation weak, better replaced by a collection of points.

Search based cluster techniques can be either deterministic or stochastic. Deterministic techniques guarentee an optimal partition by performing exhaustive enumeration. Stochastic serach techniques generate near optimal partitions reasonaly quickly and guarentee asymptotic convergence to optimal partition.

Clustering is subjective by nature. Subjectivity is usually incorporated into some phase of clustering, whether it be in selection of a pattern representation, choosing a similarity measure, or cluster prepresentation. The incorporation of domain knowledge consists of ad-hoc approaches with little in common.

Clustering of large data sets is computationaly demanding and many clustering algorithms do not scale adequately. The emerging discipline of data mining has spurred developments and optimizations in this area. Clustering is used in the data mining process for segmentation of databases into homogeneous groups, predictive modelling, and visualization. If the data set is too large to fit in main memory, techniques like *divide-and-conquer*, incremental clustering, and parallel algorithm implementations have been used.

The paper review several application domains in which clustering has been successfully employed: image recognition, data mining,

[5] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 6/1 2010.