A Hierarchical Document Clustering Environment Based on the Induced Bisecting k-Means

F. Archetti[1,2], P. Campanelli[1,2], E. Fersini[1], and E. Messina[1]

[1] DISCO, Università degli Studi di Milano Bicocca,
Via Bicocca degli Arcimboldi, 8
20126 Milano, Italy
{fersini, messina}@disco.unimib.it
[2] Consorzio Milano Ricerche, Via Cicognara 7,
20129 Milano, Italy
{archetti, campanelli}@milanoricerche.it

Abstract. The steady increase of information on WWW, digital library, portal, database and local intranet, gave rise to the development of several methods to help user in Information Retrieval, information organization and browsing. Clustering algorithms are of crucial importance when there are no labels associated to textual information or documents. The aim of clustering algorithms, in the text mining domain, is to group documents concerning with the same topic into the same cluster, producing a flat or hierarchical structure of clusters. In this paper we present a Knowledge Discovery System for document processing and clustering. The clustering algorithm implemented in this system, called Induced Bisecting k-Means, outperforms the Standard Bisecting k-Means and is particularly suitable for on line applications when computational efficiency is a crucial aspect.

1 Introduction

Document search results are often presented to user as a flat list of documents, ranked by their relevancies to a given query, and users have to examine all the titles and snippets of the documents in the list. This is a time consuming process because multiple topics can be mixed together. The need of improving the browsability of search engine results has increased the interest in different clustering approaches most of which based on vector space document models, also known as bag-of-words models [11]. As far as unsupervised classification algorithms are concerned, several approaches have been proposed [5][19] which play an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. An interesting clustering approach has been proposed in [18], which provides a mechanism to group those documents whose snippets share similar phrases, by using suffix trees. This approach is well suited for clustering web search; the only drawback is that the clustering is flat, while a hierarchical structure is usually more suitable for browsing on line search results, in particular when queries are about general topics possibly belonging to

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 257 – 269, 2006.

different domains. Clustering algorithms that build meaningful hierarchies out of large document collections are ideal tools for their interactive visualization and exploration as they provide data-views that are consistent, predictable, and at different levels of granularity. The Scatter/Gather system [9], is an interactive environment which supports the browsing, of both summaries and contents, of all the texts in a collection. It uses a hierarchical agglomerative clustering algorithm, known as Buckshot [3], which is a combination of two approaches: k-Means and hierarchical agglomerative clustering (HAC). HAC works by considering each data point as a separate cluster and then combining them in new clusters. This continues until there are only k clusters and finally, the centroids of the clusters are used as the initial centroids for the k-Means algorithm. Another important clustering system for web search results is proposed in WebACE Project [6], which uses an algorithm, named Principal Direction Divisive Partitioning [1]. It constructs a binary tree hierarchy of clusters by encompassing the entire document collection, and recursively split clusters on the bases of a linear discriminant function derived from the principal direction, until a desired number of clusters are reached. Recently, commercial products as Vivisimo (http://vivisimo.com) and iBoogie (http://www.iboogie.com/) are available on the web. They are meta-search engines that add to the flat list of query result a hierarchical structure of document clusters. Another approach, for improving browsability of web documents, returned by a search engine, consists in classifying these entities with a model built on an existing taxonomy, such as Yahoo! (http://www.yahoo.com) or the Open Directory Project (http://www.dmoz.org). In order to build such taxonomy model Naïve Bayes based method can be applied: it performs a hierarchical classification and constructs distinct classifiers at the internal nodes of the taxonomy using all the document in its child node as training data [8]. The classification is then applied at every node until a leaf is reached.

In this paper, we propose a system for searching and clustering large corpus of documents by using an experimental approach for on line text processing and hierarchical clustering. The clustering algorithm we propose in this paper, called Induced Bisecting k-Means, is an extension of the Standard Bisecting k-Means [14][15] which makes it more stable with respect to noisy data and applicable to web search results. Preliminary experiments and evaluations are conducted to investigate its effectiveness. Results obtained seem to be promising both in terms of accuracy and computational time. The basic outline of this paper is as follows. Section 2 presents our methodological approach together with a brief description of the system architecture. In Section 3 the datasets and the performance measures used for the validation of our approach are described. In Section 4 a set of preliminary experimental results are presented and, finally, in Section 5 conclusions are derived.

2 System Model

We propose a Knowledge Discovery System able to satisfy the following requirements:

- *High Dimensionality*: it can process documents with thousands of relevant terms
- *Modular and Extensible*: the system is designed in a modular way, so that new functionalities could be easily added
- *Speed*: search and clustering features return relevant results in a few seconds