ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of DEVELOPMENT RESEARCH

International Journal of Development Research
Vol. 4, Issue, 12, pp. 2669-2671, December, 2014

Review Article

MINIMIZING SPACE USING VICTORIAN REPRESENTATION BY CREATING WORD CLUSTRING

***Selvi, K.**

Department of Information Technology, Apollo Engineering College, Chennai, Tamil Nadu, India

ARTICLE INFO

Article History:
Received 30th September, 2014
Received in revised form
21st October, 2014
Accepted 29th November, 2014
Published online 27th December, 2014

Key words:
Gaussian assumption,
Association rules, Cluster,
Agglomerative Approach.

ABSTRACT

For transmission of information to be reliable we establish a communication via ATC (Automatic Text Categorization) scheme. The dimensioned document representation space which is redundant turn's problematic in many criteria's. The purpose of a paper to eliminate the disadvantage by brute force association rules and Gaussian assumption which minimizes (reduces) space by compression and using Victorian representation creating cluster of word for easy search.

INTRODUCTION

An Introduction to Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Architecture for Data Mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large data warehouse.

****Corresponding author: Selvi, K.***
Department of information technology, Apollo engineering college, Chennai, Tamil Nadu, India

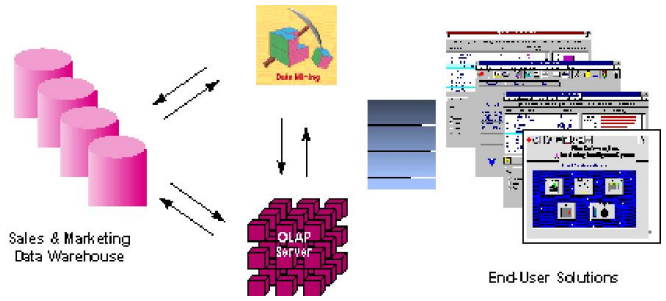

Figure 1. Integrated Data Mining Architecture

ATC machine architecture

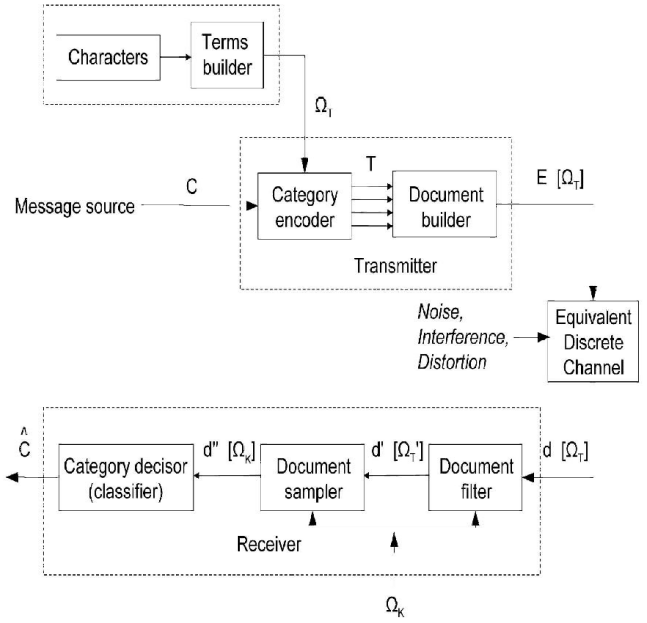

Figure 2. ATC architecture

Clustering of information retrieval

Partition unlabeled examples into disjoint subsets of clusters, such that, Examples within a cluster are very similar; Examples in different clusters are very different. Discover new categories in an unsupervised manner (no sample category labels provided).

Clustering Algorithm

Agglomerative Approach

Agglomerative (bottom-up) methods start with each example in its own cluster and iteratively combine them to form larger and larger clusters. The algorithm is simple and scales well to large vocabulary sizes, since instead of comparing the similarity of all pairs of words; it restricts the comparison to a smaller subset of size M (M being the final number of clusters desired. The Algorithm initializes the M clusters to the M first words of the sorted list. It follows on by iteratively comparing the M cluster and merging the closer ones. Empty clusters are filled with next words in the sorted list. When merging occurs, the distribution of the new cluster becomes the weighted average of the distributions of its constituent words. For instance, when merging terms t_j and t_k into a same cluster, the resulting distribution function is:

$$f_{t_j v_{t_k}}(c) = \frac{p(t_j)}{p(t_j)+p(t_k)} f_{t_j}(c) + \frac{p(t_k)}{p(t_j)+p(t_k)} f_{t_k}(c)$$

Ft-variance of discrete signal
P-probability mass function
Hierarchical Agglomerative Clustering:

Assumes a similarity function for determining the similarity of two instances. Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster. The history of merging forms a binary tree or hierarchy. Start with all instances in their own cluster. Until there is only one cluster:

Among the current clusters, determine the two clusters, c_i and c_j that are most similar. Replace c_i and c_j with a single cluster $c_i \cup c_j$.

Cluster Similarity

- Assume a similarity function that determines the similarity of two instances: *sim*(*x*,*y*).

Cosine similarity of document vectors.

- How to compute similarity of two clusters each possibly containing multiple instances?

Single Link: Similarity of two most similar members.
Complete Link: Similarity of two least similar members.
Group Average: Average similarity between members

Single Link Agglomerative Clustering

- Use maximum similarity of pairs:
- Can result in "straggly" (long and thin) clusters due to *chaining effect*.

$$sim(C_i, C_j) = \max_{x \in C_i, y \in C_j} sim(x, y)$$

Complete Link Agglomerative Clustering

- Use minimum similarity of pairs
-

$$sim(C_i, C_j) = \min_{x \in C_i, y \in C_j} sim(x, y)$$

Victorian Representation in text categorization

We study an approach to text categorization that combines distributional clustering of words and a Support Vector Machine (SVM) classifier. This word-cluster representation is computed using the recently introduced Information Bottleneck method, which generates a compact and efficient representation of documents. When combined with the classification power of the SVM, this method yields high performance in text categorization. This novel combination of SVM with word-cluster representation is compared with SVM-based categorization using the simpler bag-of-words (BOW) representation. The comparison is performed over three known datasets. On one of these datasets (the 20 Newsgroups) the method based on word clusters significantly

outperforms the word-based representation in terms of categorization accuracy or representation efficiency.

Conclusion

The theoretical model we have proposed has led to a performing two-level term-space reduction scheme, implemented by a noisy term filtering and a subsequent redundant term compression. We are currently pursuing our work with the design of a divisive clustering algorithm, which, in view of the results obtained with the tested agglomerative clustering schemes, we think can throw interesting improvements both in classification effectiveness and computational efficiency terms. We have also envisaged establishing a thorough similarity measures comparison and analysis. Future work is also foreseen in the communication theoretical modeling aspect, with special stress on the synthesis of prototype documents via the generative model proposed, as well as the deepening on the document coding (and subsequent decoding) optimal design.

REFERENCES

[1] Marta Capdevila and Oscar W. Marquez Florez. "A Communication Perspective on Automatic Text Categorization." IEEE Transaction on Knowledge And Data Engineering, Vol.21, NO.7, July 2009

[2] Sebastiani, F. "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[3] Joachims, T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. 10th European Conf. Machine Learning (ECML), pp. 137-142, 1998.

[4] Joachims, T. Learning to Classify Text Using Support Vector Machines—Methods, Theory, and Algorithms. Kluwer/Springer, 2002.

[5] Baker, L.D. and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. Special Interest Group on Information Retrieval (SIGIR '98) 21st ACM Int'l Conf. Research and Development in Information Retrieval, pp. 96-103, 1998.

[6] Slonim, N. and N. Tishby. "The Power of Word Clusters for Text Classification," Proc. 23rd European Colloquium on Information Retrieval Research, 2001.

[7] Dhillon, I., S. Mallela, and R. Kumar. "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research (JMLR), special issue on variable and feature selection, vol. 3, pp. 1265-1287, 2003.

****