Chapter XIII

Mining Message Board Content on the World Wide Web for Organizational Information

Cheryl Aasheim
Georgia Southern University, USA

Gary J. Koehler
University of Florida, USA

ABSTRACT

This chapter proposes a methodology to scan, analyze and classify the content of primarily text-based Web documents to aid an organization in gathering information. The representation and classification of the document is based on the popular vector space model and linear discriminant analysis, respectively. The methodology is developed and demonstrated using real chat room discussions about a publicly traded company collected over a 12-day period. The purpose of this chapter is to develop and demonstrate a methodology used to aid an organization in its environmental scanning efforts, in light of the vast quantities of information available via the Internet.

INTRODUCTION

Data and knowledge derived from data is critical to an organization. The amount of data available to an organization for possible analysis via sources both internal and external is vast. Processing, interpreting, using and managing data is paramount to an organization's success. One source of information is the World Wide Web (WWW). There are many documents of primarily text-based data available on the WWW including documents from chat rooms, message boards and news documents. Content analysis of such documents can provide an organization with useful information for decision-making. One corporation, Opion, is making use of Web documents to obtain useful information for an organization by monitoring chat rooms (Wakefield, 2001).

The purpose of this paper is to provide a process to scan, analyze and classify the content of Web documents in an automated manner, in order to provide useful information to an organization for the purpose of decision-making. The Web documents analyzed are text-based messages from chat rooms or message boards that are written about various organizations. The content of message boards can be used to detect possible future changes in the stock returns or trading volume for an organization, its key vendors, distributors and/or competitors. This chapter will outline how such documents could be organized, analyzed and used by an organization to gain valuable information.

The process developed in this chapter combines many areas of research including the vector space model (VSM) introduced by Salton (1968), linear discriminant analysis introduced by Fisher (1936), environmental scanning (Aguilar, 1967) and text classification methods (for examples, see Hayes & Weinstein, 1990; Lewis & Ringuette, 1994; Apte, Damerau & Weiss, 1994; Wiener, Pedersen & Weigend, 1995; Lewis, Schapire, Callan & Papka, 1996; Cohen & Singer, 1996; Moulinier, Raskins & Ganascia, 1996; Yang, 1994; Vapnik, 1995; Cortes & Vapnik, 1995). First, a brief review of each of the areas of research is provided. Next, we describe the process used to analyze the Web documents and provide an example. The chapter concludes with a discussion of the implications of our process in practice.

LITERATURE REVIEW

Environmental scanning is a process that involves gathering and using information from an organization's environment to aide management in decision-making (Aguilar, 1967; Choo & Auster, 1993; Lester & Waters, 1989). Aguilar (1967) was the first to classify the types of search an organization uses to scan the environment. The categories he developed are undirected viewing, conditioned viewing, informal search and formal search. Choo (1995) provides a framework for examining the environmental scanning literature prior to 1995. He organizes the literature into the following research categories: information needs as the focus of environmental scanning, information seeking use and preferences, information seeking through scanning methods and information use. He reviews the major contributions in each area.

Several studies have found a link between organizational performance and scanning activity (Miller & Friesen, 1977; Newgren, Rasher & LaRoe, 1984; Dollinger, 1984; West, 1988; Daft, Sormunen & Parks, 1988; Subramanian, Fernandes & Harper, 1993;

Related Content

Hybrid Partitioning-Density Algorithm for K-Means Clustering of Distributed Data Utilizing OPTICS

Mikoaj Markiewiczand Jakub Koperwas (2019). *International Journal of Data Warehousing and Mining (pp. 1-20).*

www.irma-international.org/article/hybrid-partitioning-density-algorithm-for-k-means-clustering-of-distributed-data-utilizing-optics/237135

A Blind Signature-Based Approach for Cross-Domain Authentication in the Cloud Environment

Aniello Castiglione, Francesco Palmieri, Chin-Ling Chenand Yao-Chung Chang (2016). *International Journal of Data Warehousing and Mining (pp. 34-48).*

www.irma-international.org/article/a-blind-signature-based-approach-for-cross-domain-authentication-in-the-cloud-environment/143714

Semi-Supervised Sentiment Classification on E-Commerce Reviews Using Tripartite Graph and Clustering

Xin Lu, Donghong Gu, Haolan Zhang, Zhengxin Song, Qianhua Cai, Hongya Zhaoand Haiming Wu (2022). *International Journal of Data Warehousing and Mining (pp. 1-20).*

www.irma-international.org/article/semi-supervised-sentiment-classification-on-e-commerce-reviews-using-tripartite-graph-and-clustering/307904

Domain-Driven Data Mining: A Practical Methodology

Longbing Caoand Chengqi Zhang (2006). *International Journal of Data Warehousing and Mining (pp. 49-65).*

www.irma-international.org/article/domain-driven-data-mining/1774

Early Works in Spatio-Temporal Mining

Wynne Hsu, Mong Li Leeand Junmei Wang (2008). *Temporal and Spatio-Temporal Data Mining (pp. 112-129).*

www.irma-international.org/chapter/early-works-spatio-temporal-mining/30264