Strip-Mining the Web with SAIM:
A System for the Analysis of Instant Messaging

David G. Schwartz, Bar-Ilan University, Israel
Zac Sadan, Bar-Ilan University, Israel

ABSTRACT

There is no shortage of knowledge waiting to be mined on the Web. But there is a difference between mining to the depths of the Web and examining what is occurring on its surface. In this article, the authors argue that the search for actionable knowledge should start as close as possible to the user, which, in today's technological environment, means tapping into Instant Messages. They present a System for the Analysis of Instant Messages, explain the architectural and operational motivation, and show how instant messages can be mined to produce a new source of Web-based knowledge.

Keywords: communications; data mining; instant messages; knowledge management

INTRODUCTION

Most research on data mining the Web for actionable knowledge inherently assumes that the knowledge in question has been buried deeply. Why, otherwise, would we be required to mine for it? Digging deeply into Web-based knowledge presents significant challenges ranging from extracting and formatting data for human use (Knoblock, Lerman, Minton, & Muslea, 2000) to correlating the discovered data with the current application action (Ramakrishnan & Grama, 1999) to improving a search by mining the link structure of the Web (Chakrabarti et al., 1999) to mining data from mobile users (Goh & Taniar, 2005). A different approach to finding actionable knowledge is to tap into that knowledge as it is being created — knowledge in action, so to speak. To retain the mining analogy, this could be referred to as the information analog to strip-mining — the process of extracting ore from wide yet shallow mineral deposits. Finding knowledge, or minerals, for that matter, closer to the surface has clear economic
advantages both in terms of time and equipment costs. Fortunately, the main disadvantage to strip mining in the physical world — ecological damage — does not play a part in strip mining for knowledge.

Data mining is fundamentally an applied science that focuses on how to find useful knowledge in data (Wu et al., 2003). Taking the view that data mining should be application-driven, we stand to benefit from moving our mining activity as close as possible to that application rather than in logs or databases to be analyzed post-application. To clarify, our approach is not one of data mining; rather, we contend that an alternative to data mining must be sought in order to better utilize the vast quantities of knowledge that flow across the Web. Harnessing tacit knowledge is recognized as one of the most challenging aspects of knowledge management. Tacit knowledge embodies what the knower knows, based on experience, beliefs, and values (Marwick, 2001; Nonaka & Takeuchi, 1995). It is actionable knowledge at its finest, and the text of instant messages (IM) is the closest we can hope to get to a stream-of-consciousness in which tacit knowledge is first exposed. The ultimate aim of our approach is to enable the structured analysis of IM in real time and to enable a correlation between message content and internal corporate data. The applications of such an approach will be to link IM to the corporate operational fabric rather than have it function as a stand-alone communications medium.

We are addressing the challenge of harnessing the tacit knowledge to be found in IM. In turning our attention to knowledge closer to the surface, we have developed and are experimenting with a System for the Analysis of Instant Messages (SAIM). The activities performed by the system combine packet-sniffing techniques with a knowledge management system in order to improve the performance of IM activity based on message content and accumulated knowledge. On the one hand, capture the knowledge related to user behavior and usage patterns, and on the other hand, act on this knowledge before it gets buried in the depths of a database or log file.

Earlier work on actionable knowledge (Schwartz & Te’eni, 2000) focuses on tying knowledge to action in an e-mail environment, where e-mail messages are augmented with organizational memories to add information richness. The current research takes this in a new direction by performing real-time analysis of IM content correlated with historical message content and organizational knowledge.

IM is a Web-based technology that enables users to synchronously send and receive messages, files, and other data within seconds of transmission. Initially developed as a method to keep in touch with friends in real time, the technology underlying these functions has become increasingly important to business. IM now extends its reach from the enterprise to the home to mobile devices with always-on connections. IM offers the convenience of e-mail, combined with the immediacy of a telephone conversation and the potential for file and voicemail message transmission, but perhaps most importantly, it