

## **HotMiner: Discovering Hot Topics from Dirty Text**

**Malu Castellanos**  
**HP Labs, USA**

Abstract: For companies with web sites that contain millions of documents available to their customers, it is important to identify what are the customer's hottest information needs along with the documents satisfying these needs, and to organize this information into hot topics. In particular, customer service centers can achieve enormous savings by making hot topics readily available on their web sites. Customers will then easily find the documents that answer their questions without needing to make a call to be handled by a customer representative. This is the case with technical support centers where fewer support engineers would be needed if customers were able to efficiently find the right information to self-solve their problems. As a matter of fact, it is common that about 80% of customers' problems refer to the same kinds of problems, and therefore, by identifying the topics of these hot problems and making them directly available on the web, these customers could potentially self-solve their problems and the support staff could be greatly reduced.

Customer support centers usually have search logs and case logs that contain information related to customers' problems when they try to self-solve them or open a case to have a support engineer solve them, respectively. The text in both kinds of logs usually exhibits various kinds of dirty or noisy features introduced by the clicking and typing behavior of customers and support engineers or by the nature of the problem solving process itself. In this chapter we present a novel approach to mining these logs to discover hot topics of customers' problems. We describe a technique to mine search logs to discover not only high quality hot topics but topics that match the user's perspective as well, which often is different from the topics derived from document content categorization methods. We also describe techniques to mine case logs in order to complete the picture of the hot topics. In contrast to most text mining work, our approach deals with dirty text and includes a variety of techniques to directly solve dirty features or at least to be robust in spite of them. In particular, these techniques include: a post-filtering technique to deal with the effects of noisy click streams in query logs; a thesaurus assistant to help in the generation of a thesaurus of "dirty" variations of words and used to normalize the terminology; and a summarizer to eliminate irrelevant text from non-narrative noisy text containing programming code, special symbols, incorrect use of English grammar, cryptic tables, ambiguous and missing punctuation and use of domain-specific jargon. The techniques that compose our approach have been implemented as a toolbox, HotMiner, which has been used in experiments on logs from Hewlett-Packard's Customer Support Center