

Paper 358-2009

Web Crawling and Search as a First Step to Textual Data Gathering

Yves Schabes, Teragram Corporation

ABSTRACT

Valuable textual data comes from sources that are more and more diverse. This includes well-controlled document repository systems but also, to an increasing degree, external Web pages. Web pages such as blogs and Wiki entries have very dynamic content which can be difficult to track. However, this content is more and more critical to the company's ability to make the right business decisions (reacting to customer complaints on a given product, for instance).

An elaborate Web Crawling solution is needed to track such content on a large scale and with a low latency. We will discuss how such a solution, combined with Search capabilities provides the means to import this textual data efficiently.

No paper was submitted for publication.

CONTACT INFORMATION

Yves Schabes
Teragram Corporation
Yves.Schabes@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.