Automatic reproduce metadata from the log of HTTP server

Toshihiro Aoyama[1], Yuta Suzuki[1], Kazutsuna Yamaji[2]

1. Suzuka National College of Technology
2. National Institute of Informatics

Many academic organizations have been establishing institutional repository (IR) to gather digital publications. Most access of the IR is from search engines, e.g. Google, Bing and Yahoo. On the other hand, the search performance of the repository itself is not well. One of the reasons is that the major search engines crawl PDF files of IRs in order to index keywords in a PDF resource, although most of the repository systems do not handle a full text of PDF resources. It is necessary to enrich metadata of resources of the IR in order to increase the search performance. However there is a limit to register a large amount of metadata manually.

When a user, who wants to search bibliographies or references, clicks a resource in a repository, a query used the search is related to the resource in the user criteria. In other words, the user labels the resource with the search query. The accessed resource and the query is stored in an access log of the HTTP server.

Therefore we propose that a new metadata can be collected from access logs of a HTTP server. After a access log from search engines is picked out, a search query and an accessed resource id is extracted from the log. If the query is not included in a metadata of the id, the query is candidate for new metadata. To verify whether it is suitable to use the search query as new resource metadata, we calculated the concordance rates between the keywords extracted from HTTP log and metadata of the accessed resource.

The result showed that 60.1% of search keywords did not correspond to metadata. Detail analysis also showed that most of not matched keywords (83.7%) are the words that are included in the full text of the resource and the remaining (16.8%) is related to the resource. Therefore, the search performance can be improved by appending new metadata, which used as search query. A new keyword extracted from access log of HTTP server will be submitted to a repository server using SWORD protocol.

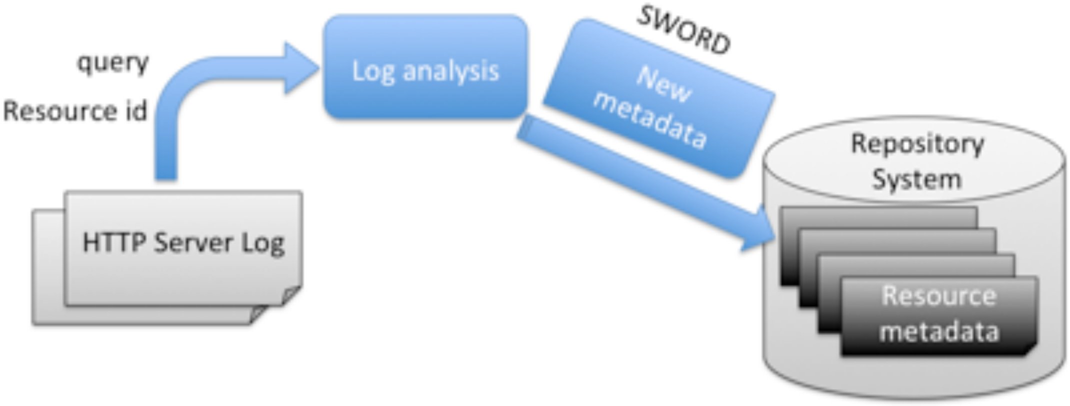

Figure 1. System architecture of proposed system