論文 / 著書情報
Article / Book Information

Title	A Method for Searching Keyword-lacking Files Based on Interfile Relationships
Author	Tetsutaro Watanabe, Takashi Kobayashi, Haruo Yokota
Journal/Book name	Proc. of 16th Intl Conf. Cooperative Information Systems(CoopIS2008), , , pp. 14-15
発行日 / Issue date	2008, 11
DOI	10.1007/978-3-540-88875-8_7
権利情報 / Copyright	The original publication is available at www.springerlink.com.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

A Method for Searching Keyword-lacking Files Based on Interfile Relationships

Tetsutaro Watanabe[1], Takashi Kobayashi[2], and Haruo Yokota[1]

[1] Grad. School of Information Science and Engineering, Tokyo Institute of Technology, Japan
{tetsu@de., yokota@}cs.titech.ac.jp
[2] Grad. School of Information Science, Nagoya University, Japan
tkobaya@is.nagoya-u.ac.jp

Abstract. Traditional full-text searches cannot search keyword-lacking files, even if the files are related to the keywords. In this paper, we propose a method for searching keyword-lacking files named FRIDAL (File Retrieval by Interfile relationships Derived from Access Logs). The proposed method derives interfile relationship information from file access logs in the file server, based on the concept that those files opened by a user in a particular time period are related.

1 Introduction

Advances in information technologies have led to many types of multimedia data being stored as files in computer systems alongside conventional textual material. Moreover, the recent price drop for magnetic disk drives has accelerated the explosive increase in the number of files within typical file systems [1]. To find a desired file located at a deep node in the directory tree, several desktop search tools using full-text search techniques have been developed. However, their target is restricted to text-based files such as Office documents, PDFs, and emails. Other types of files, such as image files and data files, cannot be found by these full-text search tools because they lack search keywords. Even for text-based files, they cannot be found if they do not include directly related keywords. It becomes even harder if these files are located in different directories from the files that contain the keywords.

To address the demand for searching for these keyword-lacking files, we focus on the relationship between files that have been frequently accessed at about the same time. Although several researches for deriving interfile relationship from system call/OS event logs have been proposed[2, 3], these methods need to modify OS of target systems and/or to install custom plugins and did not consider detail access patterns of target files.

In this paper, we propose a method for mining the file access logs in a file server to find interfile relationships and for searching keyword-lacking files that match with given keywords by using interfile relationships.

2 Proposed Method

First, we extract FUD (File Use Duration) of each files as the time between open-file and close-file from the file access logs. However, the actual duration of file use differs from the FUD because several applications do not keep file opened while using it and/or a user sometimes leaves his or her seat with files open.

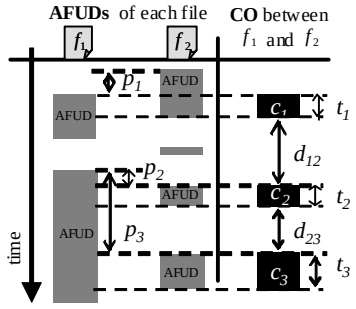

Fig. 1. Calculation of relationship elements

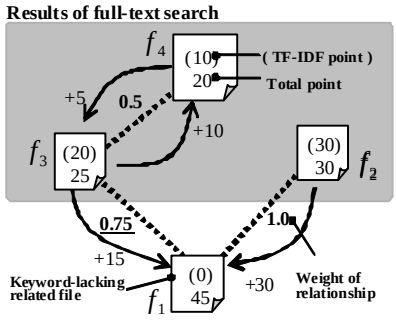

Fig.2. Calculation of the point of files.

In our proposed method, we prepare the *File Type List* to indicate which application keep file opened and calculate the *Active Time List* from the file access log. By using above two information, we extract *AFUD*(Approximate FUD) of each files.

We assume that strongly related files are used at the same time when executing the same task. To express this relationship, we introduce four "relationship elements", where the term "CO" (co-occurrence) is defined as the overlap of two AFUDs; T (Total time of COs), C (Number of COs), D (Total time of the time span between COs) and P (Similarity of the timings of the open-file operations). Fig 1. shows how we calculate each elements. By using these four relationship elements, we define the weight of interfile relationship as follows: $R(f_i, f_j) = T^{\alpha} \cdot C^{\beta} \cdot D^{\gamma} \cdot P^{\delta}$. Finally we calculate file point of f_i by adding tf.idf point of f_j in proportion to the normalized $R(f_i, f_j)$ (Fig.2).

3 Conclusion

This paper presents a method for searching for files that lack keywords but do have an association with them. The proposed method derives interfile relationship by extracting AFUD of each files and calculating four "relationship elements" from AFUDs.

Although we cannot describe details due to limitations of space, we have implemented the proposed method FRIDAL as an experimental system. It can mines the interfile relationships from the access logs of Samba and performs the file point calculations by using interfile relationships and a full-text search engine, Hyper Estraier.

We also have evaluated its effectiveness by experiments. We have compared the search results for FRIDAL with a full-text search method, directory search method and a method used in Connections[2]. The experiment showed FRIDAL is superior to other methods in the 11-points precision and the recall/precision of the top 20.

References

1. N. Agrawal, W. J. Bolosky, J. R. Douceur and J. R. Lorch: A Five-Year Study of File-System Metadata. ACM Trans. on Storage Vol.3, Iss.3, No.9 (2007)
2. C. A. N. Soules, and G. R. Ganger: Connections: using context to enhance file search. Proc. SOSP'05, pp.119-132 (2005)
3. R. Ohsawa, K. Takashio, H. Tokuda: OreDesk: A Tool for Retrieving Data History Based on User Operations, Proc. ISM'06, pp. 762-765 (2006)