Structured Ontology and Information Retrieval for Email Search and Discovery

Peter Eklund and Richard Cole

School of Information Technology and Electrical Engineering
The University of Queensland
St. Lucia, QLD 4072
peklund@itee.uq.edu.au, rcole@itee.uq.edu.au

Abstract. This paper discusses an document discovery tool based on formal concept analysis. The program allows users to navigate email using a visual lattice metaphor rather than a tree. It implements a virtual file structure over email where files and entire directories can appear in multiple positions. The content and shape of the lattice formed by the conceptual ontology can assist in email discovery. The system described provides more flexibility in retrieving stored emails than what is normally available in email clients. The paper discusses how conceptual ontologies can leverage traditional document retrieval systems.

1 Introduction

Client-side email management systems are document management systems that store email as a tree structure in analog to the physical directory/file structure. This has the advantage that trees are simply explained as a direct mapping from the structure of the file system to the email. The disadvantage is that at the moment of storing an email the user must anticipate the way she will later retrieve the email. How then should email be organized? Should we store it as a specialization or a generalization hierarchy? Are we trying to give every email a unique key based on its content or cluster emails broadly on their content?

This problem generalizes to other document types, organization is both context and query dependent (after the fact). One such organization of an *associative store* is a virtual file structure that maps the physical file structure to a view based on content. Information retrieval gives us the ability to index every meaningful word in a text by generating an the inverted file index. The index can then be reduced by stemming, compression and frequency analysis. These scalable techniques from information retrieval can be extended by re-using conceptual ontologies as a virtual file structure.

In this paper, we profile *HierMail*[1] (previously referred to in various stages as Cem, ECA or Warp9) that follows from earlier work in medical document retrieval reported in [3]. HierMail is a lattice-based email retrieval and storage program that aids in knowledge discovery by a conceptual and virtual view over

[1] see *http://www.hiermail.com*

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 75–84, 2002.

email. It uses a conceptual ontology as a data structure for storing emails rather than a tree. In turn, formal concept analysis can be used to generate a concept lattice of the file structure. This permits clients to retrieve emails along different paths and discover interesting associations between email content.

In HierMail, email retrieval is independent of the physical organization of the file system. This idea is not new, for instance, the concept of a *virtual folder* was introduced in a program called View Mail (VM)[5]. A virtual folder is a collection of email documents retrieved in response to a query. The virtual folder concept has more recently been popularized by a number of open-source projects[2]. Other commercial discovery tools for email are also available, see *http://80-20.com* for example. HierMail differs from those systems in the understanding of the underlying structure – via formal concept analysis – as well as in the details of implementation. It therefore extends the virtual file system idea into document discovery.

Concept lattices are defined in the mathematical theory of *Formal Concept Analysis* [4]. A concept lattice is derived from a binary relation which assigns attributes to objects. In our application, the objects are all emails stored by the system, and the attributes *classifiers* like 'conferences', 'administration' or 'teaching'. We call the string matching regular expressions *classifiers* since HierMail is designed to accommodate any form of pattern matching algorithm against text, images or multimedia content. The idea of automatically learning classifiers from documents has been the focus of the machine learning and text classification communities[1] but is not specifically considered in this treatment.

2 Background

Formal Concept Analysis (FCA) [4] is a long standing data analysis technique. Two software tools, Toscana [7] and Anaconda embody a methodology for data-analysis based on FCA. A Java-based open-source variant of these programs, called TOSCANAJ, has also been developed[3]. Following the FCA methodology, data is organized as a table in a RDBMS and modeled mathematically as a multi-valued context, (G, M, W, I) where G is a set of objects, M is a set of attributes, W is a set of attribute values and I is a relation between G, M, and W such that if (g, m, w_1) and (g, m, w_2) then $w_1 = w_2$. In the RDBMS there is one row for each object, one column for each attribute, and each cell can contain an attribute value. Organization over the data is achieved via conceptual scales that map attribute values to new attributes and are represented by a mathematical entity called a *formal context*.

A *conceptual scale* is defined for a particular attribute of the multi-valued context: if $\mathbb{S}_m = (G_m, M_m, I_m)$ is a conceptual scale of $m \in M$ then we require $W_m \subseteq G_m$. The conceptual scale can be used to produce a summary of data in the multi-valued context as a *derived context*. The context derived by $\mathbb{S}_m = (G_m, M_m, I_m)$ w.r.t. to plain scaling from data stored in the multi-valued context

[2] see *http://gmail.linuxpower.org/*
[3] see *http://toscanaj.sourceforge.net*