MICE$^3$: An Information Desktop on the Web

Gan Keng Hoon, Saravadee Sae Tan and Bryan Gan
Comdev Software Sdn. Bhd.
{gan | sara}@mice3.com and cdgan@comdevweb.com

Abstract
Resembling the desktop of personal computer in which files and applications are stored, MICE$^3$ desktop is a web-based information desktop where individual can keep and manage resources such as links to web sites and services, as well as web sources and documents. Resources on the desktop are represented in web directory structure and annotated with information such as concepts, entities, etc. Categories are used to organize resources in the directory. Each category is annotated with concept learnt from the contents of the category.

MICE$^3$ desktop allows users to exploit information and knowledge from their resources. Users can retrieve information by navigating the directory structure, look up the resource index or searching the resource contents. The desktop also allows users to effectively manage their resources by enabling them to enhance categories with concept. The concept is further used to assist user in organizing resources into their directory.

1. Introduction

The availability of web search technologies like Google and Lexxe allow us to discover endless resources on the web. Everyday, we find valuable resources relevant to our needs, such as files (e.g. document, audio, video, photo), services (e.g. banking, news, entertainment sites), all kinds of web sites etc. In order to well utilize resources that we have accumulated, there is a need to have an intelligent system that helps us in managing and retrieving these resources efficiently.

The idea of MICE$^3$ desktop is to provide an environment on the web for individual to exploit information and knowledge for their daily activities. We target on handling resources in a few aspects: (i) representation of personal resources using web directory structure, (ii) annotation of resources with concepts, (iii) classification of resources into categories, and (iv) concept-based and contextual-based resource retrieval.

The functionalities of MICE$^3$ desktop is presented in Section 2 and relevant research are discussed in Section 3. In Section 4, we show the MICE$^3$ architecture model that supports the design of our information desktop.

2. The MICE$^3$ Desktop

2.1. The MICE$^3$ Resources

We currently support two groups of resources, i.e. links (e.g. portals, sites, services, etc) and documents (e.g. html, txt, etc). Links on MICE$^3$ desktop resemble Google’s Bookmark or Microsoft IE’s Favorites; whereas documents include text files in our local desktop.

2.2. Resource Retrieval

Often, individual has large amount of resource collections but lack of useful methods to access and look up them. This leads to low utilization of available resources due to difficulty in finding them though they exist in personal storage.

MICE$^3$ desktop adopts the web directory concept for navigation purpose. Different from web directory, we allow organization and annotation of individual resources. Resources in the directory are categorized and annotated with useful information extracted from the resource contents such as entities, concepts, links, email etc. User can navigate the directory structure and refer to the resource summary to look up desired information.

User can also look up available resources using the index on the desktop. The index is a set of keywords representing the subjects or topics of information on the desktop.

MICE$^3$ desktop provides a more meaningful search function by allowing user to include concept-based keywords in a query. For example, the query “book: machine learning” will return information about ‘machine learning’ which is relevant to the concept ‘book’, instead of ‘course’, ‘conference’ etc.

2.3. Resource Management

Just like real life documents that require filing effort, desktop resources require proper management effort, which is to create proper categories and to organize the resources into correct categories. However, when too many categories are created, individual tends to forget the purpose of the categories and also the contents that they have stored inside. Hence, problems such as overlapping of categories and poor utilization of available contents often happen.
To effectively manage and organize the resources, every category in the directory is enhanced with concept to indicate the purpose or topic of the category [4]. User can manually annotate the category with concept, or request the system to learn the concept from the contents of the category. To keep a new resource in the directory, the user can request the system to suggest suitable categories.

3. The MICE³ Research

On MICE³ desktop, we employ a few research methods to support a more effective resources retrieval and management.

3.1. Learning Category Concept

The understanding or knowledge about a category as well as its underlying contents is made explicit through the usage of category concept. The concept reflects the semantic and the purpose of the category created. A learning approach is used to obtain the concept of a category. Basically, concept is a set of features (descriptive keywords) used to annotate a category and can be exploited for information retrieval, resources look up and knowledge discovery. Concept features are selected automatically from the contents of the category or specified manually by user.

Our category concept learning process involves two phases, i.e. training texts selection and features selection. First, training texts are selected from the contents of the category. The selection of training texts is crucial as only significant contents can reflect the concept of the category. At this stage, we employ a simple structural parsing approach to extract text segments as our training texts. Then, features are selected from the training texts. Our feature selection algorithm generates a set of candidate features using suffix array structure [3] and weights the candidates based on their frequency and structural information. An optimum set of candidates is selected as features.

3.2. Recognizing Resource Concept and Entity

The structure (i.e. organization of content) or markup of a resource (e.g. HTML document) can express the relationships and the importance of information segments in the resource [1]. Thus, instead of interpreting the contents of our desktop resources as ‘flat’ texts, we utilize the structural information to determine the boundaries of text segments in order to identify entities and their concepts. The concept-based entity is used as lexical knowledge to facilitate our resource retrieval task [2].

3.3. Mapping Query-Resource

Resources on the desktop are annotated with metadata which can be extracted from the process of concept recognition, acquired from structured data such as XML, WSDL or manually specified by domain expert. We interpret the meaning of a resource using the semantics learnt from the metadata. These semantics are used as knowledge in the searching process to obtain answers that are more relevant to the search query.

4. The MICE³ Architecture Model

The architectural design of MICE³ is shown in Fig 1.

5. The Demonstration

In this demonstration, we will show a resource directory created on our MICE³ desktop. The directory consists of links to web sources and documents. We will show the learning of categories concepts in the directory. User is allowed to interact with the look up of resources and classification of new resources.

6. References