

FOCI : Flexible Organizer for Competitive Intelligence

Hwee-Leng Ong, Ah-Hwee Tan, Jamie Ng, Hong Pan, Qiu-Xiang Li

Kent Ridge Digital Labs

21 Heng Mui Keng Terrace

Singapore 119613, Republic of Singapore

{hweeleng, ahhwee, jamie, panhong, qiuxiang}@krdl.org.sg

ABSTRACT

This paper describes how an integrated web-based application, code-named FOCI (Flexible Organizer for Competitive Intelligence), can help the knowledge worker in the gathering, organizing, tracking, and dissemination of competitive intelligence or knowledge bases on the web. It shows how text mining techniques including a novel user-configurable clustering, trend analysis and visualization techniques can be used synergistically to address the problem of managing information gathered from the web. FOCI allows a user to define and personalize the organization of the information clusters according to their needs and preferences into portfolios. Predefined sections for organizing information in specific domains is also supported. The personalized portfolios created can be saved and subsequently tracked and shared with other users. In addition, FOCI is designed to handle multilingual documents.

Keywords

Competitive intelligence, text mining, web mining, clustering, personalization, trend analysis, visualization

1. INTRODUCTION

More company and news information are increasingly available on the web. As such, it has become a gold mine of on-line information that is crucial for competitive intelligence (CI). To harness this information, various search engines and text mining techniques have been developed to gather and organize it. However, the user has no control on how the information is organized through these tools and the information clusters generated may not match their needs. The process of manually compiling documents according to a user's needs and preferences and into actionable reports is very labour intensive, and is greatly amplified when it needs to be updated frequently. Updates to what has been collected often require a repeated search, filtering of previously retrieved documents and re-organizing.

We present an integrated web-based application, called FOCI, that supports the human in the following key activities in

the CI cycle: gathering, organizing, tracking and dissemination of CI from the web. With FOCI, a user can perform a search to gather the documents from multiple desired sources and organize them automatically into information portfolios. These portfolios can be manipulated by the user according to their needs and preferences and subsequently saved or disseminated to other users. The portfolio allows a user to track new information.

FOCI differs from existing systems/products in its unique way of allowing users to personalize the organization of information portfolios according to their needs so as to facilitate future gathering and tracking of topics of interest from the web.

1.1 Related Works

Most current search engines (such as Yahoo!, Excite, Alta Vista, etc) retrieve information upon user's search queries but do not organize the search results. Those that organize (such as Copernics, BullsEye, NorthernLight etc) do not support creation, maintenance, and/or manipulation of information portfolios. Internet portals (such as My Yahoo! and My Catcha) offer personalized content deliveries that allow users to define profiles and automatic news or alerts based on user profiles through emails. However, they do not provide facilities to maintain and track information of specific topics. A number of established CI companies (e.g. Wincite, Correlate, and STRATEGY! etc) provides means for users to define their business landscapes for gathering relevant information. But they do not provide means for organizing and managing domain information and knowledge. Knowledge management tools (e.g. Knowledge Server, Knowledge Organizer and iMiner for Text etc) provide facilities for processing and organizing text-based information. However, they do not provide the personalization capability as needed to build and maintain personal information portfolios according to individual needs and preferences.

The rest of this paper describes the FOCI architecture (Section 2), an example scenario illustrating FOCI's capabilities (Section 3), user-configurable clustering (Section 4), and concluding remarks (Section 5).

2. FOCI System Architecture

Figure 1 shows the architecture of FOCI. It comprises an Information Gathering module for retrieving relevant information from the web sources; a Content Management module for organizing information into portfolios and personalizing the portfolios; a Content Mining module for discovering new information and a Content Publishing module for publishing and sharing of information and a user interface front end for graphical visualization and users interactions. The portfolios created are stored into CI knowledgebases which can be shared by the users within an organization.

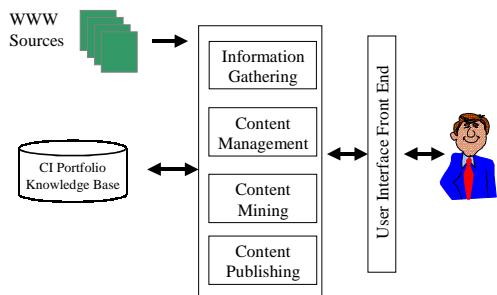


Figure 1. FOCI System Architecture

3. AN EXAMPLE SCENARIO

Here we run through a scenario to demonstrate how FOCI can aid in the key activities of the CI cycle.

3.1 Gathering

This is handled by the Information Gathering module. FOCI is implemented as a server-based application and accessed via a web browser. Once the user has logged into FOCI, he gathers information by creating a new information portfolio. He assigns a portfolio name and issues a search command by specifying the keywords and the sources that he would like to search. The top n numbers of search results from each site can also be specified.

3.2 Organizing

This is handled by the content management module that makes use of a patent-pending user-configurable clustering (UCC) technology [1] to organize and personalize the clusters.

3.2.1 Types of Organization

UCC organizes by clusters or by using a domain template. The former refers to the clustering output of typical clustering engines. The latter refers to organization into predefined sections particular to a domain and clusters appearing under each section.

The system caters for five criteria to be considered for clustering: *Title, Description, Site, Country* and *Organization*. To support real-time content aggregation and clustering, FOCI only looks at the description supplied by the search engines instead of loading the original documents. It also looks at the URL addresses which provide much meta information of web pages.

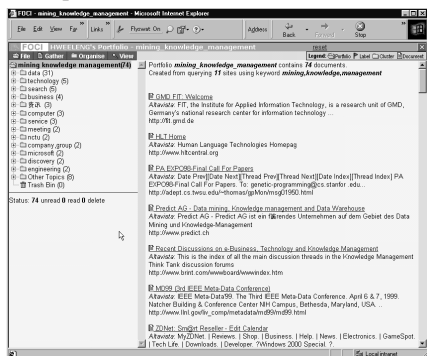


Figure 2. Organization by clusters

Figure 2 shows an example of a portfolio on the topics "mining, knowledge management" organized by clusters using criteria *Title, Description, Country* and *Organization*. The left panel shows all the clusters while the right panel shows the entire search results. Clicking on any of the clusters will display the

respective list of documents on the right panel. The cluster keywords contain a mixture of English and Chinese words as the documents searched are in both languages. Clusters with single documents are organized under a cluster called "Other Topics" to reduce the number of clusters displayed.

Figure 3 shows an example of the portfolio with same topic organized by an Information Technology (IT) domain template for CI using the same clustering criteria. This template organizes the search results into the following predefined sections: *News, Market, Companies/Products, Resources, Events* and *Others*. Within each section, the search results are organized into clusters.

In Figure 3, we find that UCC has grouped conferences that it found into a "conference" cluster under the *Events* section. This template has made the job of viewing the clusters less tedious (compared to Figure 2) in that related clusters of interest are automatically grouped together.

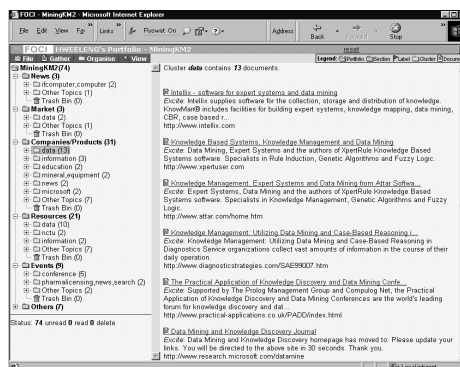


Figure 3. Organization by domain template

3.2.2 Personalizing the clusters organization

After clustering, FOCI provides a unique personalization feature for a user to customize the clusters organization according to their needs. This includes adding or deleting document links, creating, deleting, grouping or splitting folders, as well as adding and editing annotations for the portfolio, folders or documents.

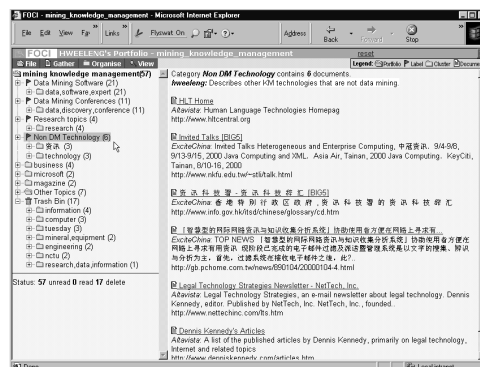


Figure 4. Partially organized portfolio

Figure 4 shows a partially organized portfolio after performing the operations of split cluster, create new cluster, group clusters, deletion and annotation on the organization in Figure 2. The big cluster "data" has been split into two groups (or labels): "data mining software", and "data mining conference". The label "Research topics" contains a user-defined cluster "research". Two of the existing clusters have been grouped under a label called "'Non DM Technology". Irrelevant clusters have

