8-2002

Organizing and personalizing intelligence gathering from the web

Hwee-Leng ONG

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Jamie NG

Pan HONG

Qiu-Xiang LI

Organizing and Personalizing Intelligence Gathering from the Web

Hwee-Leng Ong,* Ah-Hwee Tan, Jamie Ng, Hong Pan and Qiu-Xiang Li
Laboratories for Information Technology, Republic of Singapore

ABSTRACT **In this paper, we describe 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 (CI). It combines the use of a novel user-configurable clustering, trend analysis and visualization techniques to manage information gathered from the web. FOCI allows its users to define and personalize the organization of the information clusters according to their needs and preferences into portfolios. These personalized portfolios created are saved and can be subsequently tracked and shared with other users. The paper runs through an example to show how the use of a predefined domain template coupled with personalization can greatly enhance an organization and tracking of CI gathered from the web. Copyright © 2002 John Wiley & Sons, Ltd.**

INTRODUCTION

The World-Wide Web contains a wealth of company and news information waiting to be tapped. It has become one of the sources to gather competitive intelligence (CI) about other competitors or the development of new trends. To gather this information, a typical approach has been to use the search engines. However, there are some limitations to these tools. In general, they return listings of results ranked in a certain order (e.g. Google, Yahoo!, AltaVista). More sophisticated ones, such as Northern Light, BullsEye and Copernic go a step further organize the search results into folders. As in typical clustering systems (e.g. Kohonen 1988; Carpenter and Grossberg, 1987), you will realize that you have no control over how the information is organized through these tools and the information clusters generated in each folder may not match your needs. As such, these tools are used mainly for gathering purposes only. You will still need to manually compile the documents found according to your needs and preferences and into actionable reports. This process is very labor intensive, and is greatly amplified when you need the information to be updated frequently. To update, you will often do a repeated search, filter off previously retrieved documents and reorganize the information again.

According to Fuld's report (Fuld & Co., 2002), the Intelligence Cycle comprises the tasks of planning, gathering, analyzing and reporting. We have implemented a showcase system called FOCI (Flexible Organizer for Competitive Intelligence) that provides an integrated platform to facilitate the gathering, organizing, tracking and dissemination of information gathered from the web. With FOCI, you can perform a search to *gather* the documents from multiple desired sources and *organize* them automatically into a knowledge base called portfolios. These

* Correspondence to: Hwee-Leng Ong, Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613, Republic of Singapore.
E-mail: hweeleng@lit.a-star.edu.sg

International Journal of Intelligent Systems in Accounting, Finance & Management
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/isaf.208
Int. J. Intell. Sys. Acc. Fin. Mgmt. **11**, 9–21 (2002)

portfolios can be manipulated according to your needs and preferences and subsequently saved or *disseminated* to other users. In addition, the portfolio is not static but can allow you to *track* new information added to it. The above system could also be customized within an intranet environment to capture and share knowledge bases or competitive intelligence knowledge within an organization.

There are many existing tools or products that perform one or more of the above-mentioned CI activities. Most current search engines (such as Yahoo!, Excite, AltaVista, etc.) retrieve information upon a user's search queries but do not organize the search results. Those that organize (such as Copernic, BullsEye, Northern Light, Vivisimo, etc.) do not support creation, maintenance, and/or manipulation of information portfolios. One has to use the search tools to perform the search and organize periodically and maintain their information portfolios manually to keep the content up to date. There exist Internet portals (such as My Yahoo! and My Catcha) that 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. There are a number of established CI companies (e.g. Wincite, STRATEGY, etc.) that provide means for users to define their business landscapes for gathering relevant information. Some knowledge management tools (such as Knowledge Server, Knowledge Organizer, iMiner for Text, ClearResearch Suite, etc.) provide facilities for processing and organizing text-based information. However, they do not provide the type of personalization capability found in FOCI. The novelty and main differentiating aspect of FOCI from existing systems/products lies in its unique way of allowing its users to personalize the organization of information portfolios according to their needs in order to facilitate future gathering and tracking of topics of interest from the web.

We will now cover the FOCI architecture as well as run through an example scenario to illustrate how you can use FOCI to create your own CI knowledge base. In addition, we also provide a description of the underlying technology behind FOCI, followed by a conclusion.

FOCI SYSTEM ARCHITECTURE

Figure 1 shows the architecture of FOCI. It comprises:

- an Information Gathering module for retrieving relevant information from web sources;
- a Content Management module for organizing information into portfolios and personalizing the portfolios;
- a Content Mining module for discovering new information;

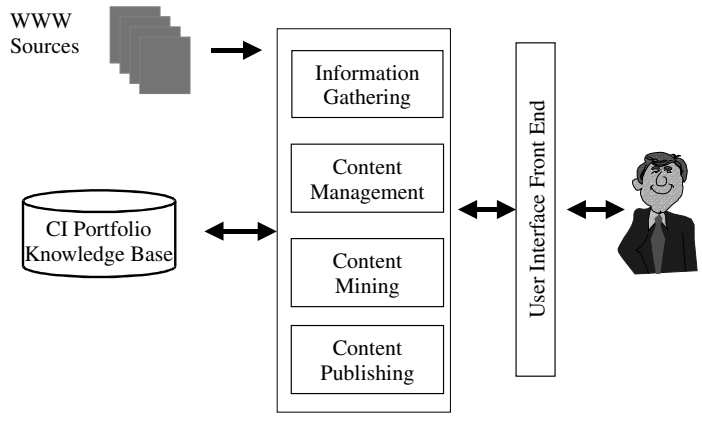

Figure 1 FOCI system architecture

- a Content Publishing module for publishing and sharing of information; and
- a user interface front end for graphical visualization and user interaction.

The portfolios created are stored into CI knowledge bases that can be shared by its users within an organization.

Information Gathering Module

This module allows users to build an information portfolio by searching and integrating information given by major search engines and news sites. The architecture is flexible to support integration of specific search, news sites or subscribed databases depending on an organization's needs. You can also insert your own links or documents not found in the search results into the portfolio directly. An automatic tracking function monitors a selected set of online sources and updates the portfolio with new content at periodic intervals.

Content Management Module

This module provides utilities to organize and manage your CI in your preferred way. There are currently two common methods to organize a document collection: clustering and categorization. Clustering organizes information automatically into groups based on similarity functions and thresholds. For categorization, the information has to be first organized into predefined sets or classes by a human before being trained by the program. There are pros and cons to both approaches. The former is able to identify new themes automatically but it gives the user no control over how the groupings appear. Further, when new documents are added to the collection, the themes or documents represented in a cluster may change. The latter approach requires effort to train the system and gives good control over the classes but lacks flexibility in being able to handle new classes. In FOCI, we have combined the two approaches into a patent-pending technology called User-Configurable Clustering (UCC) to organize and personalize the clusters. Through UCC, the user can perform functions like labeling, adding, deleting, grouping and splitting of clusters. Additional functions built on top of UCC for FOCI include annotation portfolio creation, and deletion functions.

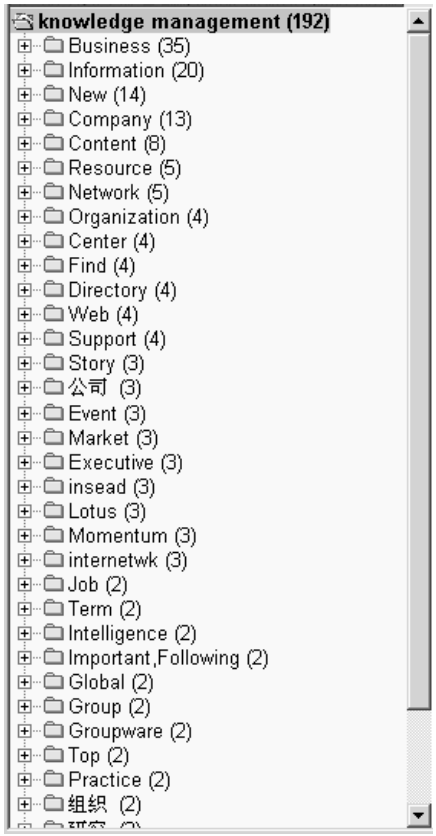

Figure 2 Organization by clusters

UCC offers two ways of organizing the information. The first, as shown in Figure 2, is by clusters, an output that is typical of clustering systems. Here, we organize a search on the top 25 results on 'knowledge management' by clusters, from six search sites and two news sites, from the largest to the smallest clusters. The cluster keywords contain a mixture of English and Chinese words as the documents searched are in both languages. Clusters with single document are organized into a cluster called *Other Topics* to reduce the number of clusters displayed. However, in this organization, many clusters are generated with a variety of topics. One way of managing this large number of clusters is through the use of domain-specific templates to group them into

several useful categories, the second way UCC organizes. Figure 3 shows a sample Information Technology domain template, which organizes into six predefined sections: *News*, *Market*, *Companies/Products*, *Resources*, *Events* and *Others*. Within each section, the search results are organized into clusters. Most of the search results from news sites are grouped under the *News* section. The *Market* section tries to identify those results that may potentially contain some market figures or reports. *Companies/Products* contains companies or product information; *Resource* tries to identify information sources; *Events* tries to identify events; while *Others* contains the rest that do not fit into the other sections. Figure 3 shows two snapshots of the expanded view for all the sections except *Others* for the same search on '*knowledge management*'.

In clustering, UCC provides five criteria to consider: *Title*, *Description* (as described by the search engines), *Site*, *Country* and *Organization* (from which the document is retrieved). The examples in Figures 2 and 3 use the *Title*, *Description*, and *Organization* criteria to derive the clusters. To support the real-time content aggregation and clustering, FOCI only looks at the information supplied by the search engines instead of loading the original documents. It also looks at the URL addresses, which provide much meta information of the web pages.

After the initial clustering, the UCC allows you a unique way to personalize the clusters that

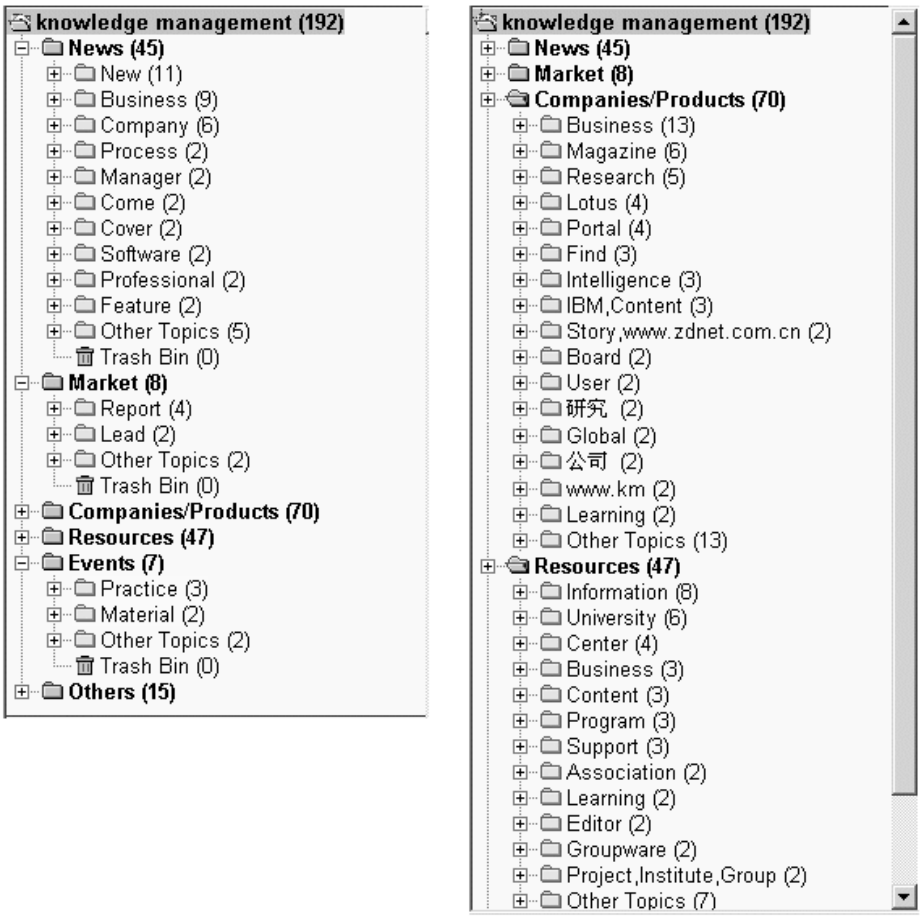

Figure 3 Organization by IT domain template

current clustering technologies do not provide. This means that you can reorganize the clusters in your preferred way. You can label clusters or a document (in this case, a search result), create new clusters that are not found, delete clusters or search results that are not of interest, remove cluster keywords that are too general by redistributing, split clusters, group clusters, and add new URLs to the portfolio. UCC will reorganize the clusters accordingly after taking into account your indicated preferences.

Content Mining Module

This module extracts key attributes from raw documents and transforms them into intuitive forms for knowledge discovery. Key analysis functions include trend analysis (finding out how a topic changes over time), topic detection/tracking (that is, detecting emerging or hot topics), and link association (finding relationships between topics).

Content Publishing Module

This module handles permission control of individual portfolios so that you can share your portfolio with other users of FOCI or export your portfolios into HTML documents. Various views for presenting the portfolio in different levels of details are supported.

AN EXAMPLE SCENARIO

In this section we will run through scenario steps to demonstrate how FOCI can aid in the CI cycle of gathering, organizing, tracking and dissemination.

Let's say, for example, that you would like to gather some competitive intelligence on the field of 'knowledge management' (KM). You would like to find out who are the players, what are the KM developments and events, what are the KM resources available, as well as to keep track of them.

Gathering

FOCI is implemented as a server-based application and is accessed via a web browser. Users login to access the system. Once logged in, you can gather information by creating a new information portfolio. Figure 4 shows the assignment of 'knowledge management' to the portfolio name and the search field.

Next, you select the sources to search from. We have implemented some popular free search sites and news sites. However, the architecture does allow for customization to easily incorporate other relevant sites depending on users' needs. You can also specify to get the top *n* numbers of search results from each site and the timeout. In Figure 4, we will also choose the clustering criteria *Title*, *Description*, and *Organization* to organize the clusters and check the *Domain Template* option. Clicking on *Create Portfolio* will give the result as shown in Figure 3.

Organizing and Personalizing the Clusters

This section illustrates how UCC technology organizes and personalizes information portfolios in FOCI.

If we look at Figure 3, there appear to be some clusters such as *Lotus*, *Portal* and *IBM* that look relevant in the *Companies/Products* sections. A click on the *Magazine* cluster shows companies publishing KM-related magazines. *Research* shows companies doing KM research. The system has also found some *Reports* that may be useful under the *Market* section. Note that there are no clusters on *Knowledge* or *Management* as these words would be found in almost all the documents and are therefore not meaningful. There are also some clusters that do not reflect very useful words like *Come*, *Lead*, *Find* that are not applicable to this topic we are studying. This set of clusters represents an initial view of the main topics found by the clustering engine. Let's perform some personalization on these clusters to provide a more useful view of the search results.

First, we can use the *Redistribute* function to remove all the cluster words that are not meaningful. For this portfolio, we find that the clusters do not change significantly and the system did not generate new clusters. In other portfolios we have tried, the system may sometimes generate new interesting clusters. This

Figure 4 Interface for *Create Portfolio*

function is useful in telling the system not to generate such clusters in future.

Next, we represent our interest by creating labels to group or create new clusters. Creating labels is a way of retaining clusters of interest so that the system would not remove these cluster topics when it reorganizes. It is also a way for you to specify your preferred topics of interest. As the documents are labeled, new clusters may be formed and those not labeled may be removed or changed as the system reorganizes.

The *Create Cluster* function comes in useful to draw out some words we are looking for that the clustering system may not have shown, for example, *Consulting* or specific company names. It is also useful to pull out all the examples that contain a specified word. For example, using this function for the *IBM* cluster has pulled two more results instead of the previous three results. Figure 5 gives a snapshot of the portfolio after some labeling and creation of new clusters in the *Companies/Products* section. It shows three major groupings represented by clusters.

Once you have labeled the results, you might want to add your own comments for the labels or cluster or on top of the summary from the search results. This annotation is useful to take note of what you have read as well as serving as a guide to others when you disseminate your portfolios. Figure 6 shows a sample annotation of a cluster.

Examining the contents of the three groupings, you may find that there are some irrelevant results, or you may like to move the results. These can be easily done through the *Delete* or *Move* functions respectively.

Sometimes the search results may not return some sites that you would like to include in your portfolio. FOCI allows a user to add document links of sites that has not been retrieved by any of the search engines. In this scenario, we only collected the top 25 results from each search engines, so a result you are looking for may have been left out. The new document added is incrementally clustered into one of the cluster folders or a new cluster folder may be created. In

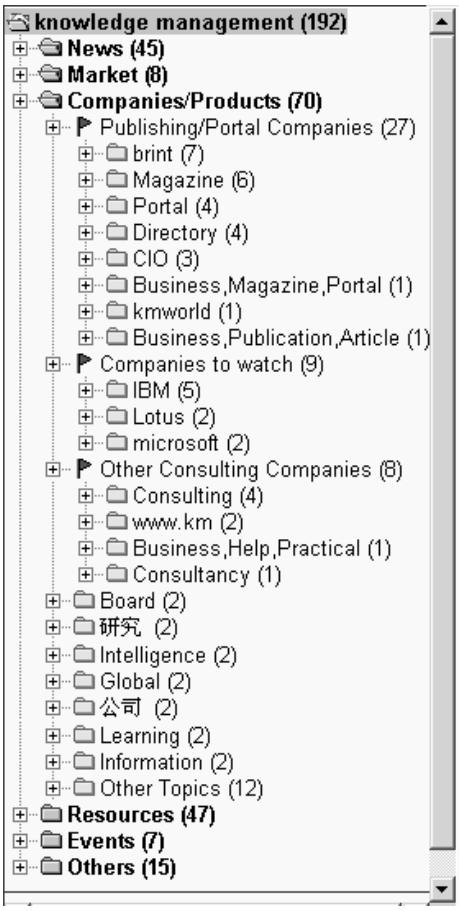

Figure 5 Partial snapshot of a section after some labeling

Figure 7 we show how we can add a URL, and description to the label *Companies to watch*.

Figure 8 shows a personalized view of our portfolio after further labeling, redistribution, addition and deletions to the other sections. Some of the created clusters (e.g. *launch*, *alliance* and *merger*) have no results but we would like future additions to this portfolio to consider these clusters as well. This view is now useful for tracking of future documents that fit the clusters defined.

Tracking

Tracking is achieved through the incremental clustering supported by the user-configurable

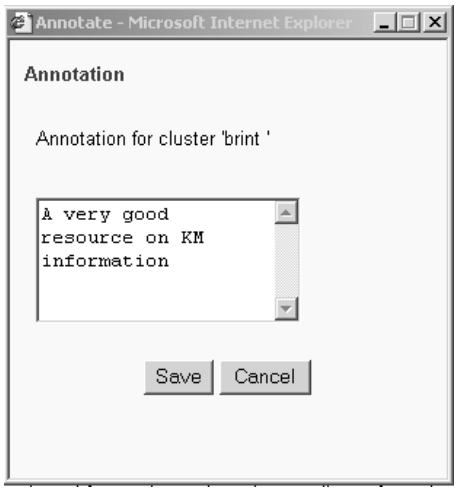

Figure 6 Example of an annotation of a cluster

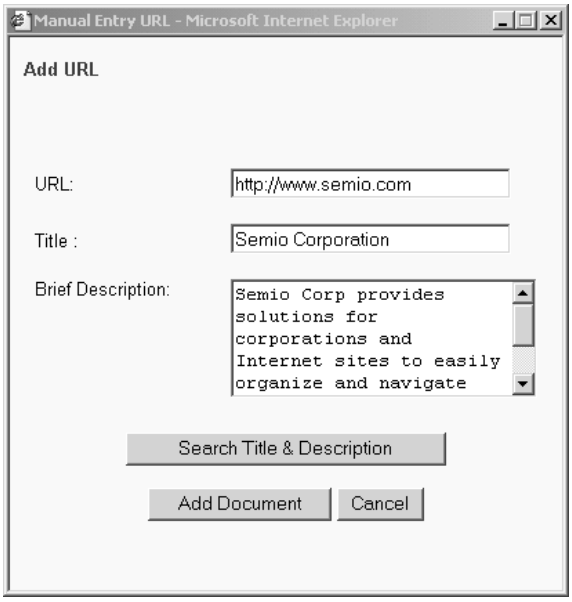

Figure 7 Adding a site

clustering system. Through FOCI, the user can specify a periodic update of the portfolio created, e.g. on a daily, weekly, bi-weekly or monthly basis. New documents found are added automatically to the collection and organized according to the folders or labels that the user has defined for

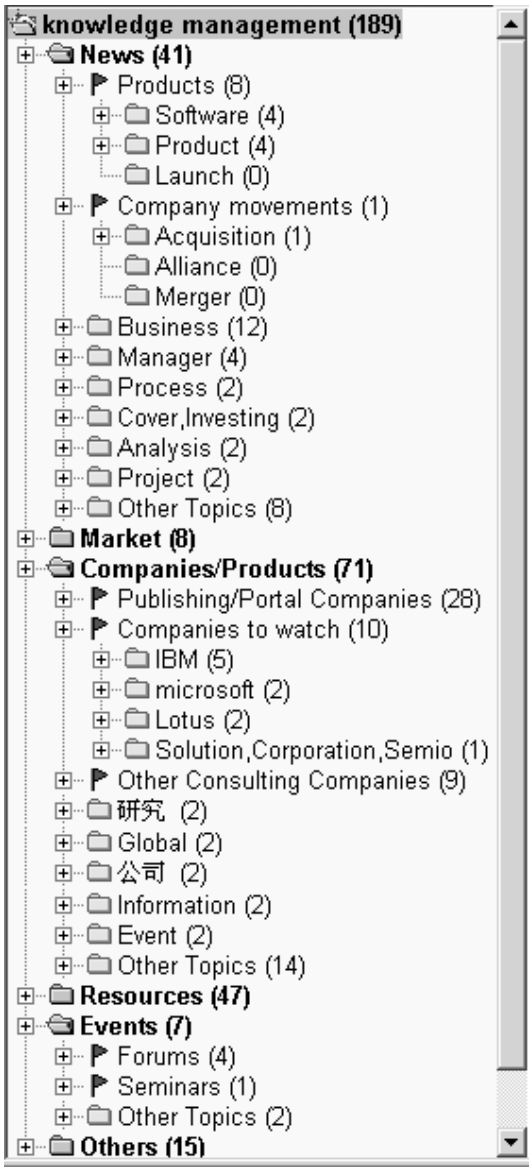

Figure 8 A personalized portfolio

that portfolio. Duplicate documents are removed. Those similar to previously deleted documents will also be removed automatically if the user had specified during the deletion so that he does not need to repeat his filtering process again. In addition, new clusters may appear that are not within the user-defined cluster structure for that portfolio. This may represent potentially new or

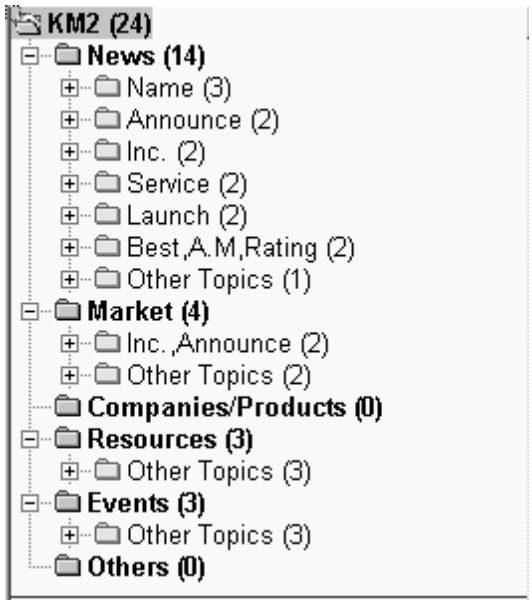

Figure 9 Clusters created based on 24 new documents without personalization

interesting information. As an example, a new set of 24 documents is collected through another news search engine. Without personalization of the portfolio, the clusters are as displayed in Figure 9.

Figure 10 shows how this new information is organized into the personalized portfolio of Figure 8. Notice that it has added three search results to the *Products* label and distributed one into *Software* cluster and two into the *Launch* cluster. There was also one further entry in the *Acquisition* cluster. The *Acquisition* and *Software* clusters were not found in the unpersonalized view (Figure 9). In Figure 10 we also discover a new cluster called *Name* under the *News* section and it describes news of appointment of people in a company. It is a cluster we had not thought of tracking before but is important to track. We can choose to move this cluster under *Company movements* to track future similar cases. Similarly, we might want to move the *Service* cluster under the *Products* label since it describes new offerings of a company. In the *Events* section of the personalized view we also discover that one more forum and one more seminar has been found

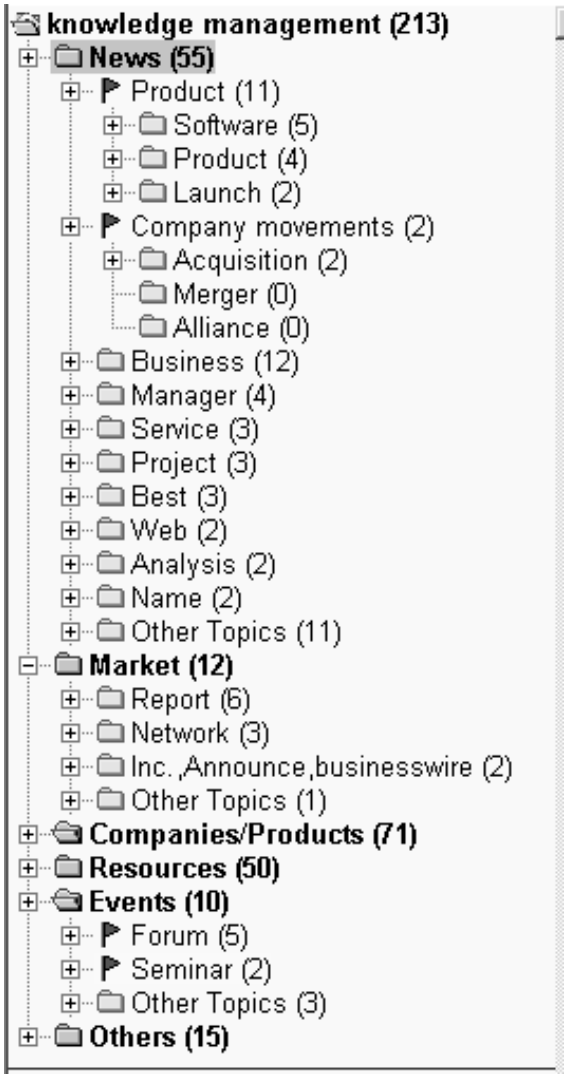

Figure 10 Organization of the new documents into the personalized portfolio

which is not so obvious in the unpersonalized view.

Dissemination

This task is supported through the Content Publishing module. Once a portfolio has been created and personalized, you can save the portfolio and share it with other users so that the knowledge captured is not lost. FOCI provides a tab for you to access all the shared portfolios. In addition, you can also publish a portfolio to a HTML document so that it can be emailed or saved. Figure 11 shows part of the HTML document generated for all the labeled clusters. User annotation for clusters or document is also captured.

USER CONFIGURABLE CLUSTERING (UCC)

In this section, we describe the underlying clustering engine that makes the above organization and personalization possible. UCC is a clustering engine based on fuzzy Adaptive Resonance Associative Maps (ARAM) (Tan, 1995; Tan and Soon, 1996) that can perform a combination of unsupervised learning (e.g. clustering) and supervised learning (e.g. categorization). ARAM belongs to a family of predictive self-organizing neural networks known as predictive Adaptive Resonance Theory (ART) that performs incremental supervised learning of recognition categories (pattern classes) and multidimensional maps of patterns. An ARAM can be visualized as two overlapping ART (Carpenter and Grossberg, 1987) modules consisting of two input fields F_1^a and F_1^b with an F_2 category field (Figure 12). Fuzzy ART (Carpenter *et al.*, 1991) which categorizes both binary and analog patterns is used here.

For each document d, we derive an information vector $\mathbf{A} = (a_1, a_2, \ldots, a_M)$ such that

$$a_i = tf(w_i)^* r(w_i)(1 - r(w_i))$$

where the term frequency $tf(w_i)$ is the number of times the keyword w_i appears in document d and the document ratio $r(w_i)$ is computed by

$$r(w_i) = \frac{df(w_i)}{N}$$

where the document frequency $df(w_i)$ denotes the number of documents in which w_i appears and N is the number of documents in the collection.

User preferences are represented by preference vectors that indicate the preferred groupings of

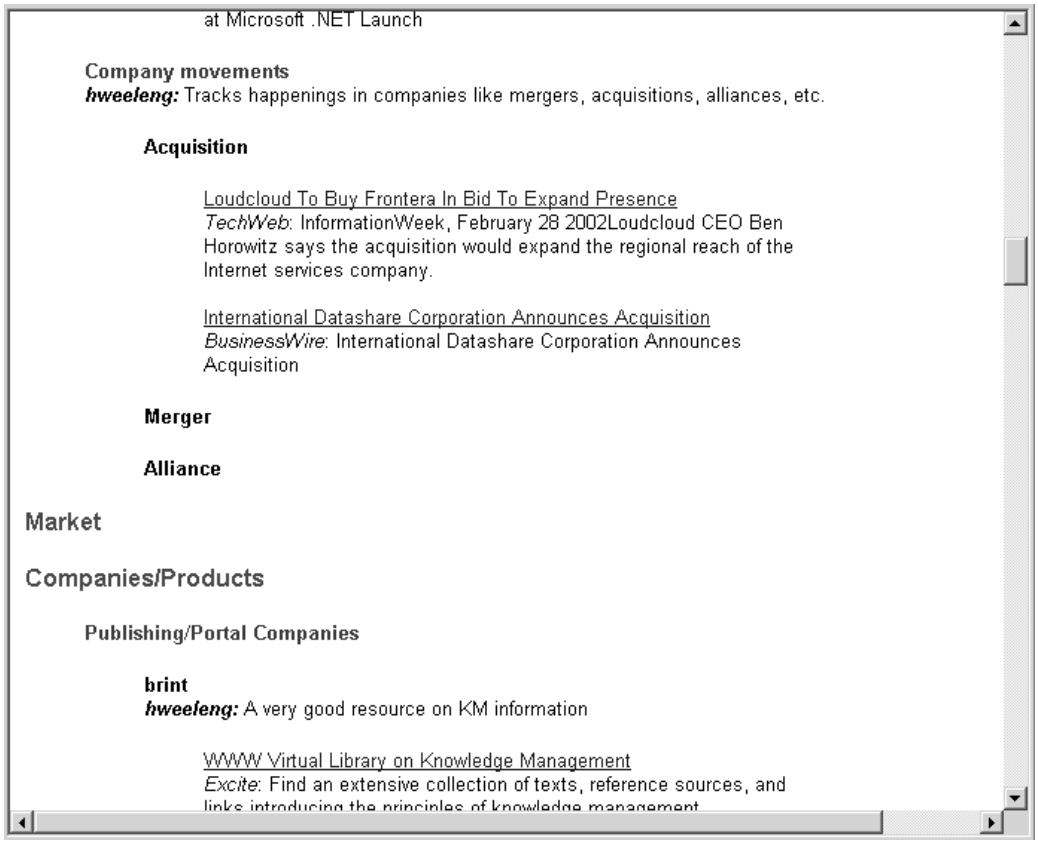

Figure 11 Publish to a HTML document

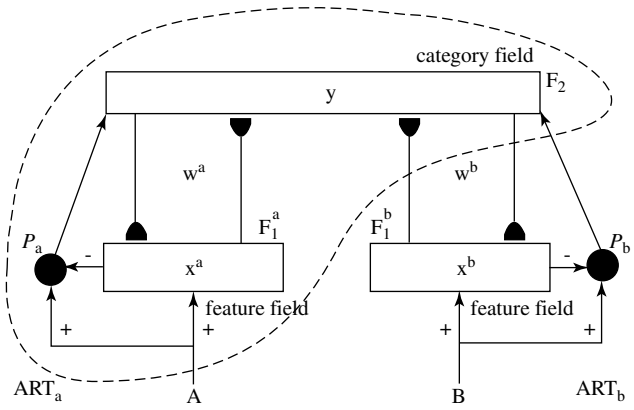

Figure 12 The Adaptive Resonance Associative Map architecture

the information. A preference vector **B** is defined by $\mathbf{B} = (b_1, b_2, \ldots, b_N)$ where b_i is either zero or one, indicating the presence or absence of the user-defined label L_i.

For user-configurable clustering, the F_1^a contains the activities of the information vectors and the F_1^b field contains the activities of the preference vectors. Information clusters (represented at F_2) are created during *learning* through the synchronized clustering of the information and preference vectors. Specifically, each recognition node or cluster j learns to encode a pair of template information vector w_j and template preference vector w_j.

To perform clustering, fuzzy ARAM operates in the typical *learning* mode to obtain assignments of information vectors. However, if a predefined cluster structure exists, the system loads the ARAM network before clustering and organizes the information according to the cluster structure. Without a predefined network structure, the system behaves like a pure clustering system that self-organizes the information based on similarities among the information vectors only.

To support the organization of predefined sections and clusters, a set of rules is applied to filter the search results into the respective sections first, followed by clustering of each section.

For personalization to occur, ARAM operates in an *insertion* mode whereby a pair of information and preference vectors can be inserted directly into the ARAM network. This mode enables a user to influence the clusters created by ARAM through indicating his or her own preferences in the form of preference vectors. For illustration, two key personalization functions are described here. More personalization functions or experimental results are found in Tan *et al*., (2001):

- *Labeling Information Clusters*: A label L is assigned to a cluster j by modifying the template preference vector w_j^b to equal **B**, where **B** is a preference vector representing L.
- *Inserting Information Clusters*: A pair of information and preference vectors (**A**, **B**) is first derived based on the key attributes of the information in the new cluster and the cluster label. During cluster insertion, fuzzy ARAM's vigilance parameters ρ_a and ρ_b are each set to 1 to ensure that only identical attributes are grouped into one recognition category. After insertion, ARAM regenerates the cluster structures by clustering all the information vectors again. Note that new clusters may be generated during reclustering.

CONCLUSION

An alpha version of the FOCI system with user-configurable clustering is available at http://textmining.krdl.org.sg/FOCI. Currently it is accessible through Internet Explorer 5.5 as the user interface is based on servlets and dynamic HTML. The FOCI server currently runs on Unix Solaris workstations. We have also implemented a topic detection and tracking component (part of the Content Mining module) that provides the user with a view of emerging or hot topics from IT news sites like Cnet and ZDNet. The emerging or hot topics can be visualized via trend graphs to see how a topic has evolved over a period of time.

As part of FOCI, it provides an integrated platform for monitoring emerging or hot topics of interest. Figure 13 shows a screenshot of the topic detection and tracking component (Kanagasa and Tan, 2001) on the FOCI login page. It shows the cluster keywords for emerging or hot topics together with representative document titles for each cluster. The user could also switch to a visual trend graph of each cluster that illustrates how that particular cluster of topic has evolved over the past few weeks.

We have presented how FOCI, an integrated system, can facilitate the gathering, organizing, tracking and dissemination of competitive information gathered from the Internet. This is done through creating information portfolios, making use of domain-specific templates and giving users control to organize them according to their preferences. New information can be discovered through its tracking functions. Also, the information portfolios can be published and shared with other users. Future directions for FOCI include:

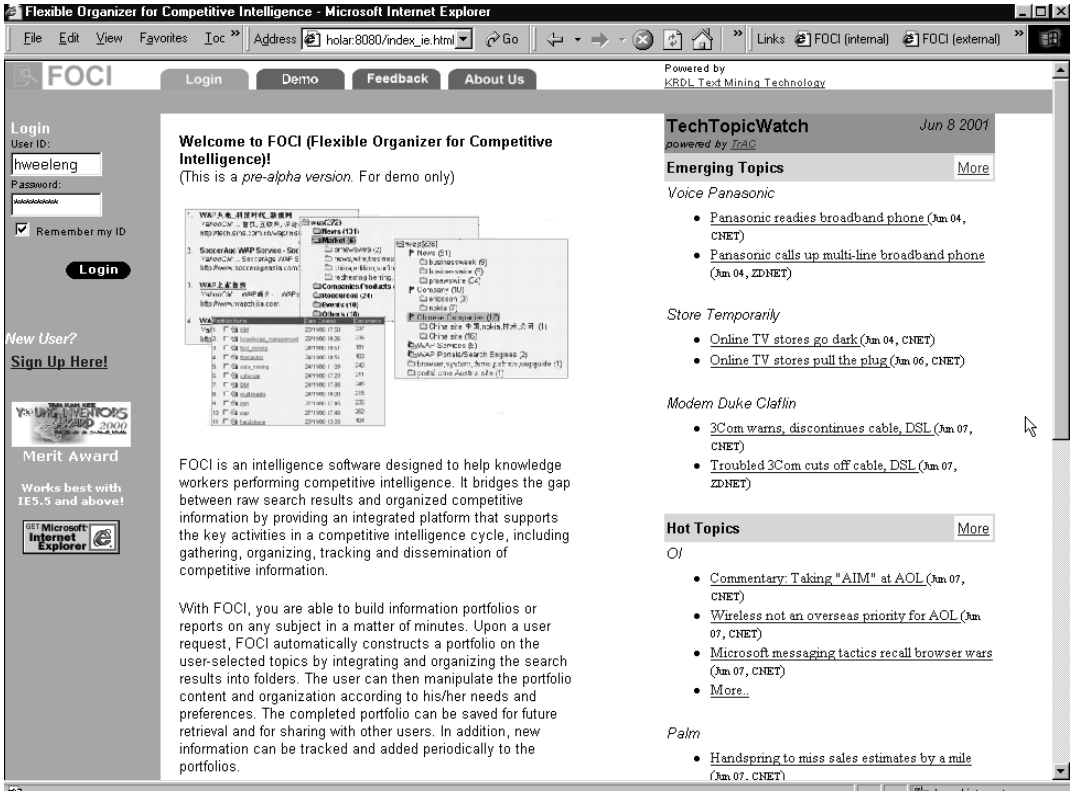

Figure 13 Topic detection / tracking component

- Improving on the results from the user-configurable clustering algorithm such as being able to handle terms (e.g. 'knowledge discovery' instead of 'knowledge' or 'discovery');
- Incorporation of other content mining functions such as link association and cluster map visualization;
- Support of other domain-specific templates;
- Support of other information sources such as on-line databases;
- Improving on the user interface through more usability testing.

ACKNOWLEDGMENTS

The contributions of Jian Su, Guo-Dong Zhou and Tong-Guan Tey are gratefully acknowledged.

References

Carpenter GA, Grossberg S. 1987. A massively parallel architecture for a self-organising neural pattern recognition machine. *Computer Vision, Graphics and Image* **37**: 54–115.

Carpenter GA, Grossberg S, Rosen DB. 1991. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks* **4**: 759–771.

Fuld & Co. 2002. A review of Software and Information Technology Offerings in the Competitive Intelligence Arena. *Intelligence Software Report 2002. Intelligence Software: The Global Evolution* (11 March 2002).

Kanagasa R, Tan AH. 2001. Topic detection, tracking and trend analysis using self-organizing neural networks. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer-Verlag: LNAI 2035, Hong Kong; 102–107.

Kohonen T. 1988. *Self-organization and Associative Memory*. Springer-Verlag: New York.

Ong HL, Tan AH, Ng J, Pan H, Li QX. 2001. FOCI: Flexible Organizer for Competitive Intelligence.

In *Proceedings of 10th ACM Conference On Information and Knowledge Processing*, Atlanta, 523–525.

Tan AH. 1995. Adaptive resonance associative map. *Neural Networks* **8**(3): 437–446.

Tan AH, Ong HL, Pan H, Ng J, Li QX. 2001. FOCI: A personalized web intelligence system. In *IJCAI workshop on Intelligent Techniques for Web Personalisation*. Seattle, 14–19.

Tan AH, Soon HS. 1996. Concept hierarchy memory model: A neural architecture for conceptual knowledge representation, learning and commonsense reasoning. *International Journal of Neural Systems* **7**(3): 305–319.