11-2001

FOCI: Flexible organizer for competitive intelligence

Hwee-Leng ONG

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

Jamie NG

Hong PAN

Qiu-Xiang LI

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.

CIKM'01, November 5-10, 2001, Atlanta, Georgia, USA.

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-contigurable 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.

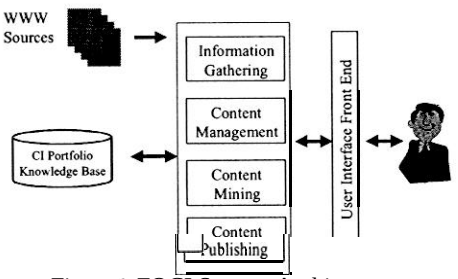

Figure 1. FOCI System Architecture

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 he considered for clustering: *Title*, Description, Sire, *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.

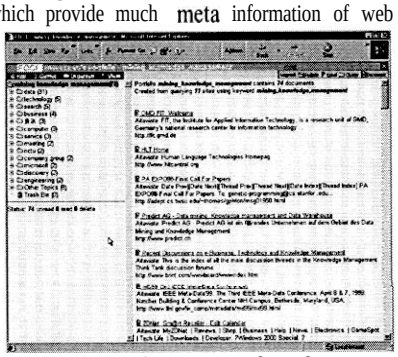
Figure 2. Organization by clusters

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, Mark,, *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.

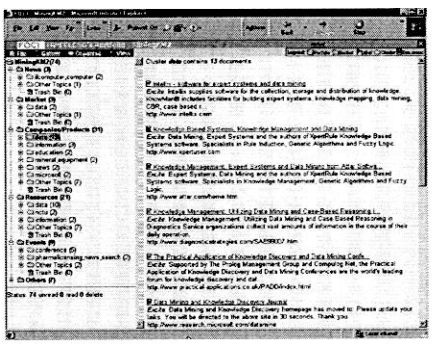
Figure 3. Organization by domain template

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.

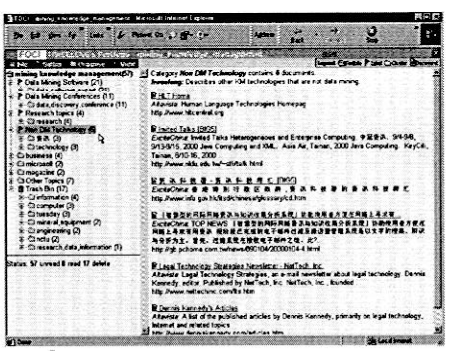
Figure 4. Partially organized portfolio

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

been moved to the trash bin to indicate to the system to delete such similar documents in future. The highlighted label "Non DM Terminology" contains annotation (at top of right panel), that may be useful for future sharing of this portfolio with other users.

3.3 Tracking

This is achieved through incremental clustering supported by UCC. The user can specify a periodic update of the portfolio, e.g. daily, weekly, bi-weekly or monthly. New documents found are added automatically to the collection and organized according to the folders or labels that the user has defined. Duplicates are removed. Those similar to previously deleted ones in the trash bin will also be deleted so that the user need not 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 interesting information.

3.4 Dissemination

This is supported by the Content Publishing module. Once a portfolio has been created, it can be saved and shared it with other FOCI users so that the knowledge captured is not lost.

4. USER CONFIGURABLE CLUSTERING

The clustering engine is based on fuzzy Adaptive Resonance Associative Map (ARAM) [2] [3] that can perform a combination of unsupervised learning and supervised learning. 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 [4] modules consisting of two input fields F_1^a and F_1^b with an F_2 category field (Figure 5). Fuzzy ART [5] is used here.

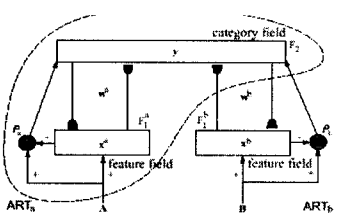

Figure 5. ARAM Architecture

For each document d, an information vector $\mathbf{A} = (a_1, a_2, \ldots, a_M)$ is derived such that $a_i = (1 + tf(w_i)) * idf(w_i)$, where the term frequency $tf(w_i)$ is the number of times the keyword w_i appears in document d and the inverse document frequency $idf(w_i)$ is computed by

$$idf = \log \frac{N}{df(w_i)}$$

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

User preferences, represented by preference vectors, indicate the preferred groupings of 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 UCC, F_1^a and F_1^b contains the activities of the information and preference vectors respectively. Information clusters (represented at F_2) are created during learning through the synchronized clustering of the information and preference vectors. Specifically, each 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. For personalization, ARAM operates in an insertion mode whereby a pair of information and preference vectors is inserted directly into the ARAM network. This enables a user to influence the clusters created by ARAM through indicating his own preferences in the form of preference vectors. Details of personalization functions and experimental results are found in [1].

5. CONCLUSION

A pre-alpha version of the FOCI system with UCC is available at http://textmining.krdl.org.sg/FOCI. It is now accessible through Internet Explorer clients only as the user interface is based on servlets and dynamic HTML. The FOCI server runs on Unix Solaris workstations. We plan to integrate a topic detection and tracking component [6], that provides the user with a view of emerging or hot topics from IT news sites like CNet. These topics can be visualized via trend graphs to see how a topic has evolved over a period of time.

Some future directions for FOCI include: improvements to UCC to include terms rather than just keywords; adding other content mining functions like link association and cluster map visualization; support of other domain-specific templates and information sources, usability testings; and reporting features.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Tan, A.-H., Ong, H.-L., Pan, H., Ng, J., and Li, Q.-X. FOCI: A Personalized Web Intelligence System. In IJCAI workshop on Intelligent Techniques for Web Personalization (Seattle, Aug 2001), 14-19.

[2] Tan, A.-H. Adaptive Resonance Associative Map. Neural Networks, 8(3), (1995), 437-446.

131 Tan, A.-H., and Soon, H.-S. Concept hierarchy memory model: A neural architecture for conceptual knowledge representation, learning & commonsense reasoning. International Journal of Neural Systems, 7(3) (1996), 305-319.

[4] Carpenter, G. A., and Grossberg, S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing, 37 (1987), 54-115.

[5] Carpenter, G. A., Grossberg, S., and Rosen, D.B. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 4 (1991), 759-771.

[6] Kanagasa, R., and Tan, A.-H. Topic detection, tracking and trend analysis using self-organizing neural networks. In Proceedings of PAKDD'01 (Hong Kong, 2001), 102-107.