International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:9, 2008

A Knowledge-Based E-mail System Using Semantic Categorization and Rating Mechanisms

Azleena Mohd Kassim, Muhamad Rashidi A. Rahman, and Yu-N. Cheah

***Abstract*—**Knowledge-based e-mail systems focus on incorporating knowledge management approach in order to enhance the traditional e-mail systems. In this paper, we present a knowledge-based e-mail system called KS-Mail where people do not only send and receive e-mail conventionally but are also able to create a sense of knowledge flow. We introduce semantic processing on the e-mail contents by automatically assigning categories and providing links to semantically related e-mails. This is done to enrich the knowledge value of each e-mail as well as to ease the organization of the e-mails and their contents. At the application level, we have also built components like the service manager, evaluation engine and search engine to handle the e-mail processes efficiently by providing the means to share and reuse knowledge. For this purpose, we present the KS-Mail architecture, and elaborate on the details of the e-mail server and the application server. We present the ontology mapping technique used to achieve the e-mail content's categorization as well as the protocols that we have developed to handle the transactions in the e-mail system. Finally, we discuss further on the implementation of the modules presented in the KS-Mail architecture.

***Keywords*—**E-mail rating, knowledge-based system, ontology mapping, text categorization.

I. Introduction

KNOWLEDGE has become an essential resource in gaining and providing a competitive edge in many organizational-level endeavors. This has led to the emergence of a knowledge-based economy where organizations are now more sensitive towards the impacts of intellectual capital and knowledge management. Riding on the knowledge management wave, various knowledge management and knowledge engineering efforts such as knowledge acquisition, organization and sharing, are keenly explored by many organizations.

Current knowledge acquisition methods face problems such as difficulties in recognizing specific knowledge when exposed to a mixture of irrelevant data. The process of testing and refining knowledge is often very complicated and usually involves high computational effort. Incomplete knowledge and poorly defined knowledge acquisition methods can also add to the seriousness of the problem.

Looking at the current trend, a lot of knowledge is circulated via e-mails and other messaging software, turning it into a popular means to acquire knowledge. Yet, traditional e-mails and messaging protocols have their own limitations that affect their effectiveness in knowledge acquisition. For example, e-mails usually form discussion threads and these series of e-mail can be a hassle for users to manage and control.

Traditional e-mails are also kept just as an archive and there are no efficient repositories or mechanisms to reuse these e-mails for decision-making purposes. This is largely due to the lack of meaning for semantic annotations to the e-mail that can help identify relevant knowledge. Although there are forum groups that employ the semantic web approach to provide semantic search capabilities, they are mostly focused on methods that search semantically by keywords. More needs to be done to build a reliable knowledge management system.

To address the issues of current e-mail systems, this paper discusses a knowledge-based semantic e-mail system (KS-Mail) for knowledge acquisition.

The knowledge acquisition efforts are further enhanced by using ontologies, allowing knowledge in the form of metadata to be captured and semantically annotated, by employing the semantic web approach. Knowledge that is acquired is then enhanced by providing semantics and machine-readable content. By using various standard representations, markup languages (e.g. XML) and other semantic technologies (e.g. ontologies), we will be able to better organize and categorize the knowledge that have been acquired.

Besides that, KS-Mail employs a forum style in handling the e-mail communication and a rating mechanism to allow evaluation of the e-mails. Therefore, the focus will be on:

- Knowledge representation: This is to allow e-mails to become machine-understandable in order to make it easier to be accessed and reused over the web.
- Knowledge organization: Ontologies are used to map the relationships between the metadata to understand the relationships, and with the help of natural language processing, knowledge is translated into a specific representation format. These capabilities are explored and

A. Mohd. Kassim is with the School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Penang, Malaysia (phone: (+604) 653 3888 ext 4392; fax: (+604) 657 3335; e-mail: azleena@cs.usm.my).

M.R. A.Rahman is with the School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Penang, Malaysia (e-mail: rashidi@cs.usm.my).

Y.-N. Cheah is with the School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Penang, Malaysia (e-mail: yncheah@cs.usm.my).

extended to enable categorization of the e-mail content, using the semantic web approach employed in a previous SEMblog project [1].

- Knowledge sharing and reuse: We introduce an architecture that allows collaborative knowledge-sharing capabilities, allowing knowledge to be evaluated/rated, and reused from time to time.

In KS-Mail, we expand the functionalities of a previous system called KM-Mail [2], with the features mentioned earlier as well as by reconstructing the application server. The reconstruction of the application server is achieved by enhancing its components such as the service manager, evaluation engine and the search engine. We have also improved the user interface to overcome problems related to massive e-mail threads and ill-managed e-mails.

II. Related Work

McDowell, Etzioni, Halevy, and Levy [3] proposed a semantic web approach to e-mail messaging that uses RDF queries tied to some descriptive texts. In order to present the class of semantic e-mail processes, they defined logical and decision-theoretic models which automatically infer which e-mail responses are to be accepted and compute optimal message-handling policy in polynomial time respectively.

In another work, a semantically enhanced e-mail system is developed where the ontology is utilized for speech acts, and non-deterministic models are presented in order to help users to decide the actions after sending or receiving e-mails [4]. This aims to overcome information overload as well as processing delays upon responding to different messages. The system is presented by creating its own e-mail speech act ontology. It is suggested that semantic annotations to an individual e-mail message can be used to enhance a semantic e-mail, which is hoped to be used to predict reactions using its predictive model.

In an effort to overcome unsolicited e-mails (spam), and protect the privacy of e-mail addresses and individual identity, there are some efforts on semantic e-mail addressing being carried out. SEA is used to address e-mails to a semantically defined set of entities [5]. SEA works to keep track of some descriptive attributes that are semantically specified to each recipient or group and is able to change dynamically.

Warren [6] discussed the use of the semantic web for knowledge management and the benefits and significance of using ontologies in the semantic web. Warren highlighted the importance of converting the semantic web into a business process value and allowing it to be available explicitly. Warren described that the building of a knowledge management environment starts from a scenario, which utilizes certain technologies to achieve the goal. If appropriate technologies are applied to work alongside semantic web, it can serve as a knowledge management scenario that can function as a search mechanism.

Some efforts have also been carried out to explore the potentials of groupware technology [7] to handle organizational knowledge. The proposed groupware is made up of components for document management, blogging and e-mail that “enforces” knowledge management within the organization. This prototype project looks forward to enhancements in terms of the groupware’s interactivity, connectivity and personalization.

III. KS-Mail Architecture

As with the previous KM-Mail [2] system, we maintain the four-layer architecture but have simplified as well as added some new features to it (see Fig. 1). The two main modules are:

- E-mail server: This module performs semantic processing to extract relevant keywords from the e-mail text and to find relationships between e-mails.
- Application server: This module provides the basic features of an e-mail system and provides a mechanism to users to evaluate or rate e-mails as well as to search for e-mails that contain particular knowledge content.

The KS-Mail repository is also an important module in the system where it stores the e-mail, discussion threads as well as the XML file for the semantic functionalities.

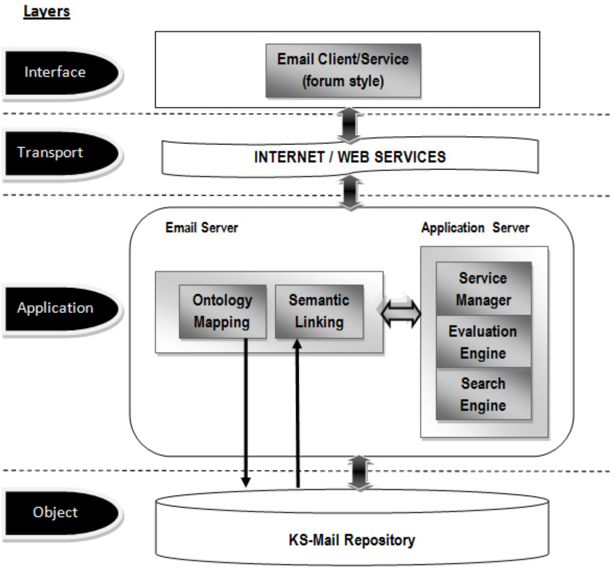


Fig. 1 KS-Mail Architecture

A. E-Mail Server

The e-mail server performs two functions: (1) ontology mapping and (2) semantic linking. As a result, this component produces categorizations of the e-mail content.

Ontology Mapping

To assist in organizing the e-mail contents as well as acquiring knowledge embedded in it, we use the Ontology Web Language (OWL), particularly OntoSem. The OntoSem ontology comes in a tangled tree of concepts, using its own metalanguage to describe the meaning of objects and events [8].

For this purpose, natural language understanding will influence the semantic representation, inference and knowledge representation during the process. In this case, word statistical distribution technique is used to categorize the e-mail contents. Subsequently, the metadata in OntoSem is mapped following its hierarchical structure.

Details of how the mapping is carried out through the ontology hierarchy in order to grab the metadata for our categorization will be discussed in Section IVA.

Ontology mapping is carried out in three main phases (see Fig. 2):

- Pre-processing: Text filtering and weight statistical distribution are carried out.
- Ontology mapping: Keywords and metadata in OntoSem are manipulated in order to produce a final list of metadata that carry different weight values.
- Storing the categories: Before the categories are finalized, sigmoid and threshold functions are applied in order to narrow down the results to more specific categories. For storage purposes, we use XML tags which will later be used to create the semantic links.

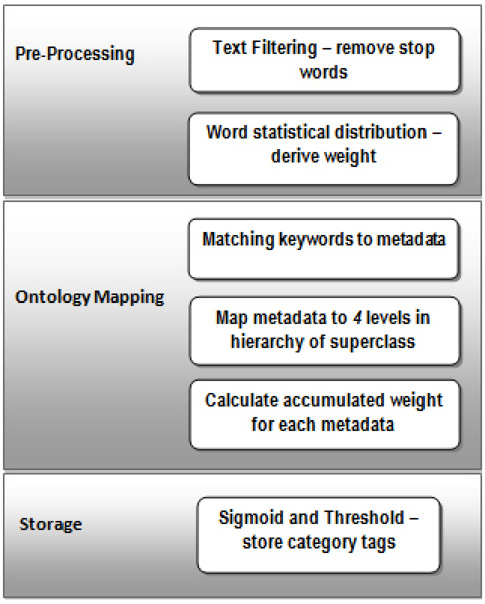

Pre-Processing
Text Filtering – remove stop words
Word statistical distribution – derive weight
Ontology Mapping
Matching keywords to metadata
Map metadata to 4 levels in hierarchy of superclass
Calculate accumulated weight for each metadata
Storage
Sigmoid and Threshold – store category tags

Fig. 2 Processes involved in the ontology mapping phase

Semantic Linking

In this component, we create semantic links between the e-mails that have same categories. Since we have a forum-styled e-mail format, we believe that the categorization will help users to manage their e-mails more effectively. Even when users are planning to evaluate or rate certain e-mails (see Section IIIB), they can locate (from their Inbox or message threads) e-mails that may be relevant to issues they want to evaluate on.

We utilize the category tags that we have generated from the ontology mapping component to enable the semantic linking. This will be further discussed in Section IVA.

B. *Application Server*

The application server has three main components: (1) service manager, (2) evaluation engine, and (3) search engine.

1) Service manager: This component handles the users' interests and activities within the KS-Mail system such as new user registration, requests to make evaluations, and semantic search. This component also provides the user interface. In short, the service manager provides the main e-mail functions as well as new features such as the rating or evaluation mechanism, and search facility.
2) Evaluation engine: This component supports the service manager by handling the rating and evaluation of e-mails. The users are able to rate an e-mail sent by another user. These ratings are sent to all recipients of that particular rated e-mail. The ratings will be averaged if more than one user rates a particular e-mail. The value will portray the quality of the e-mail content to the users.
3) Search engine: This component allows users to make certain knowledge queries. This component complements the semantic processing component carried out by the e-mail server. Using the category tags and the XML file in the KS-Mail repository, suggestions will be presented to the user according to their query.

C. *KS-Mail: E-mail Protocol*

Register and Send

If a new member wants to register into KS-Mail and join a discussion group, he/she can register through the KS-Mail server. Upon acceptance, the member is allowed to view a list of discussion groups where he/she can choose a group according to his/her interests. The discussion group's e-mail address will also be made known to the newly enrolled member.

New KS-Mail discussion groups can be added by any registered user. Upon acceptance, KS-Mail will generate a unique e-mail address for the new discussion group and this will be made known to the intended members of the group.

Receive and Reply

After the registration process, the KS-Mail server will first send an e-mail as an acknowledgement, containing instructions on utilizing the KS-Mail protocol, i.e. on how to answer the problems, evaluate answers, or reply to e-mails. The original question or problem (if any) is also included.

The recipients are allowed to issue replies to answer, or to discuss further, any problems or questions by the sender. They can do this by replying through the unique e-mail address that was created for the particular discussion group. After the reply is sent, the system will process the incoming e-mail and notify other users in the same group of the new answer or further discussion. For storage purposes, all the details of the reply, such as the message body, are parsed and extracted. As for the evaluation and searching process, all outgoing e-mail will be provided with its own KS-Mail-ID.

Read, Evaluate and Search

Each e-mail is evaluated based on its usefulness, relevance and applicability, allowing recipient to evaluate the answers or questions that appear in the discussion group.

A small form will appear in each e-mail that will allow users to rate the e-mail's content based on a scale of one (least useful) to five (most useful). Like a regular forum-based interface, a text box is provided in the e-mail where the user can input their comments regarding the answer given.

The KS-Mail repository stores all the rating or comments that are provided by the e-mail recipients. This is continuously being updated. Each rating or comment will be linked to the respective e-mail replies via the KS-Mail-ID. Fig. 3 shows the workflow for "Read and Evaluate".

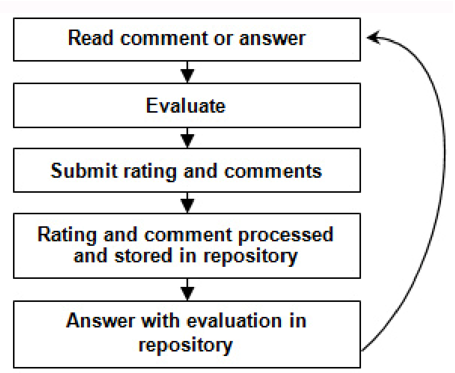

Read comment or answer
Evaluate
Submit rating and comments
Rating and comment processed and stored in repository
Answer with evaluation in repository

Fig. 3 Process flow for "Read and Evaluate"

As for the search engine, a normal keyword-based search will be used to retrieve e-mails from the archives and the knowledge repository. The categorization and linking by the semantic processing is combined with this function for a directed search. With the evaluation facility, search engine and the repository working together, it will form a tool that can be effectively used for knowledge management.

IV. Implementation

A. E-Mail Server

In this section, we provide more details on the ontology mapping and semantic linking component of the e-mail server.

Ontology Mapping

The ontology mapping technique effectively categorizes the e-mail content. The process of categorizing the e-mail content is carried out as follows:

First, we read the text from the user's e-mail entry, where all words are considered as individual keywords. Then, all stop-words are removed. When we remove stop-words (e.g. determiners, auxiliaries and preposition), we are eliminating words which are less meaningful to the content. The remaining words are then applied with the word statistical distribution to count the weight of each word.

The keywords (with their respective weights) are then mapped to the OntoSem ontology. To do this, we match each weighted word in the list to the hierarchy of classes (concepts) in the ontology. When a match is found, the probability (*P*) (see equation 1) of that class being a category is calculated.

$$P_C = D_w \, x \, L \qquad (1)$$

where *C* is the matched ontology class, *w* is the weight of the word *D*, and *L* is the level of exploration.

For each matching cycle, we will only map 4 steps up the ontology hierarchy. Therefore the value of *L* is initialized to 4. Then, the search moves up to the superclass of the earlier matched ontology class.

When a superclass is found, P_C is calculated for the matched superclass. Here, the value D_w is still the same but since the mapping is now a step higher the ontology hierarchy, the value of *L* is decreased by 1. The cycle will halt if no superclass can be found or when L = 0 (after four levels of ontology classes are mapped).

The weight for each metadata or ontology class mapped is calculated, and this is repeated for all the keywords. A class in the ontology can be mapped more than once if it is a superclass of more than one keyword. If this is the case, its weight will be a combination of all the weights.

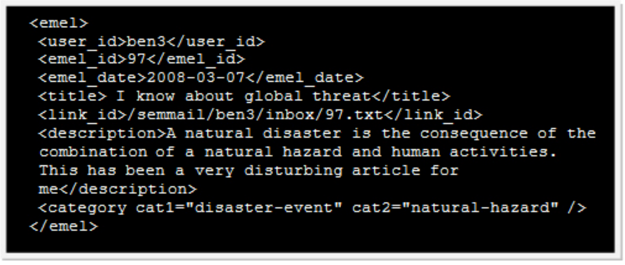
After completing the ontology mapping, a sigmoid function (see equation 2) is applied on each of the P_C values of the categories to obtain a normalized value, S_C, before listing the metadata in descending order (the higher the *Sc* value, the higher chance for that class to be chosen as a category for a particular e-mail content).

$$S_C = \frac{1}{1+e^{-(P_C)}} \qquad (2)$$

In equation 2, we set a threshold value, *T*. The categories with S_C meeting or exceeding *T* will be chosen as one of the categories for the e-mail. In this case, it is possible for an e-mail to have multiple categorizations (having more than one category) as long as their weights exceed *T*.

Finally, the categories are checked against a predefined list of categories that are "too general" to be considered as a category and are removed (these may be categories which are too high up the ontology hierarchy).

The e-mail can now be stored in the KS-Mail repository in XML format (see Fig. 4).



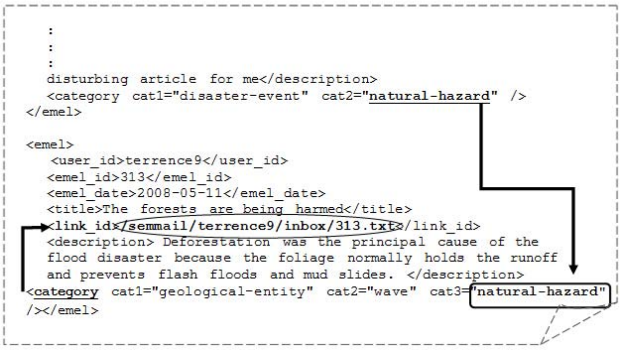
```
<emel>
 <user_id>ben3</user_id>
 <emel_id>97</emel_id>
 <emel_date>2008-03-07</emel_date>
 <title> I know about global threat</title>
 <link_id>/semmail/ben3/inbox/97.txt</link_id>
 <description>A natural disaster is the consequence of the
 combination of a natural hazard and human activities.
 This has been a very disturbing article for
 me</description>
 <category cat1="disaster-event" cat2="natural-hazard" />
</emel>
```

Fig. 4 Portion of the XML file stored in the KS-Mail repository

Semantic Linking

When the categories of the e-mail have been identified, the e-mail automatically becomes a possible semantic link to other e-mails within the same e-mail account. Therefore, when an e-mail is created or retrieved for viewing, the XML file in the KS-Mail repository is parsed through.

The e-mail server manages the search by dynamically obtaining the links contained in the `<link_id>` tags of other e-mails' XML entries that have the same categories. This is matched by comparing the `<category>` tag of the viewed/created e-mails with the existing ones (see Fig. 5).



```
   :
   :
   :
   disturbing article for me</description>
   <category cat1="disaster-event" cat2="natural-hazard" />
</emel>

<emel>
    <user_id>terrence9</user_id>
   <emel_id>313</emel_id>
   <emel_date>2008-05-11</emel_date>
   <title>The forests are being harmed</title>
   <link_id>/semmail/terrence9/inbox/313.txt</link_id>
   <description> Deforestation was the principal cause of the
   flood disaster because the foliage normally holds the runoff
   and prevents flash floods and mud slides. </description>
<category cat1="geological-entity" cat2="wave" cat3="natural-hazard"
/></emel>
```

Fig. 5 Matching categories in the XML file

When a match is found, the `<category>` tag will refer to the `<link_id>` tag to look for the identifier. The `<link_id>` stores the Unified Resource Identifier which will be used to point to the specific e-mail and be linked to it. Thus, the content of these semantically related e-mails are reused when viewed by the user.

B. Application Server

As soon as an e-mail has been sent/posted, the e-mail server will be invoked and the e-mail's text will be semantically processed. The IDs of each e-mail are also stored in the repository. Each e-mail posted by the user will be given two different IDs. The first ID is referred to KS-Mail-ID generated by the KS-Mail server, while the second ID is the reference ID that functions as a pointer to the original e-mail which the user is replying to (see Fig. 6).

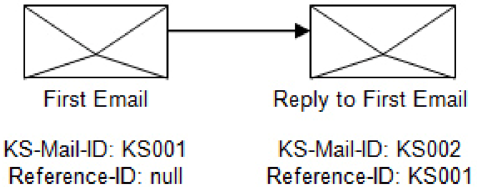


Fig. 6 KS-Mail-ID assignment mechanism

These two different IDs are important in the evaluation mechanism as each e-mail ID will be used as a reference. Whenever an evaluation process is completed, the evaluation scores in a rating form and comments will also be recorded and stored together. The process is repeated each time new evaluations come in, combining the details of the overall scores and total rating for each e-mail.

V. Conclusion

In this paper, we presented a combined approach to enhance an e-mail system through semantic categorization and rating. For this purpose, we have employed web technologies such as the semantic web, as well as ontology and natural language processing (NLP) techniques.

By using semantic technology, we allow knowledge to be acquired from the e-mail and categorized accordingly. This acquired knowledge is then stored and reused to promote knowledge sharing amongst users. The semantic link subsequently provides easy access to related e-mails within the user's account. Naturally, users cannot view or link to e-mail that is not within his/her KS-Mail account to ensure security and privacy.

What is exciting is that we have observed that KS-Mail effectively brings older e-mails to the attention of the user via the semantic links that are created. This happens whenever a new e-mail is created or received. This opens a whole new e-mailing experience for the user from a knowledge management point of view. The user is now able to carry our knowledge acquisition, organization and reuse just by using the KS-Mail system.

In summary, the value added concept of KS-Mail is the framework that highlights the importance of the e-mail's content. We also emphasize the significance of knowledge-rich e-mail through rating and searching mechanisms.

In the future, we hope to expand the features of KS-Mail to incorporate semantic blogging [7]. In this future enhancement, it is hoped that an improved ontology mapping process can be employed. Through the semantic blogging exercise, it may even be possible to tease tacit knowledge from experts – which is something that is much desired in a knowledge management context.

References

[1] A. Mohd Kassim and Y.-N. Cheah, "Using Semantic Web, Ontologies and Blogs for Knowledge Identification, Organisation and Reuse," *International Conference on Electrical Engineering and Informatics 2007 (ICEEI 2007),* Bandung, Indonesia, June 17–19, 2007, pp. 691–694.

[2] K.G. Lim and Y.-N. Cheah, "Sharing, Evaluating and Organizing E-Mail: The KM-Mail Approach," *The 3rd International Conference on Information Technology in Asia*, Sarawak, Malaysia, July 17–18, 2003, pp. 112–117.

[3] L. McDowell, O. Etzioni, A. Halevy, and H. Levy, "Semantic e-mail," *Proceedings of the 13th international conference on World Wide Web(WWW '04),* New York, NY, USA, May 2004, pp. 244–254.

[4] S. Scerri, B. Davis, and S. Handschuh, "Improving E-mail Conversation Efficiency through Semantically Enhanced E-mail," *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007),* Regensburg, Germany, September 3–7, 2007, pp. 490–494.

[5] M. Kassoff, C. Petrie, L.M. Zen, and M. Genesereth. "Semantic E-mail Addressing: Sending E-mail to People, Not Strings," *AAAI 2006 Fall Symposium*, Washington, DC, October 13–15, 2006.

[6] P. Warren, "Knowledge Management and the Semantic Web: From Scenario to Technology," *IEEE Intelligent Systems,* 2006, Vol. 21, pp. 53–59.

[7] Y.-N. Cheah, "Enhancing Groupware for Knowledge Management," *Fifth International Conference on Information Technology in Asia (CITA 2007),* Sarawak, Malaysia, July 9–12, 2007, pp. 145–150.

[8] M. McShane, S. Nirenburg and S. Beale, "An Implemented, Integrative Approach to Ontology-Based NLP and Interlingua," Working Paper 06-05. Institute for Language and Information Technologies, University of Maryland Baltimore County, USA, March 8, 2005.