Algorithms for Extracting Topic across Different Types of Documents

Shoichi Nakamura[1], Saori Chiba[1], Hirokazu Shirai[2], Hiroaki Kaminaga[1],
Setsuo Yokoyama[2], and Youzou Miyadera[2]

[1] Fukushima University, Department of Computer Science and Mathematics,
Kanayagawa 1, Fukushima, 960-1296 Japan
nakamura@sss.fukushima-u.ac.jp
[2] Tokyo Gakugei University, Division of Natural Science,
4-1-1, Nukui-Kita, Koganei, Tokyo, 148-8501 Japan
{miyadera,yokoyama}@u-gakugei.ac.jp

Abstract. Clever management of the various types of documents used in intelligent activities and their efficient utilization are important. However, most available methods target only a single type of document (e-mails, Web pages, etc.). A more promising approach is topic-centered document management. Algorithms are described for extracting topics across various of types of documents. Moreover, a topic-centered document management system is described that is based on grouping by topics.

1 Introduction

Efficient discovery and utilization of useful information from various types of documents is important in intelligent activities, e.g., research activities and cooperative software development in a network environment. However, the number of documents related to such activities increases exponentially as the activities progress. This makes it harder and harder to identify the useful documents. There is thus a strong need for ways to support the management and utilization of documents in accordance with the how the information is to be used.

Although there has been research on topic extraction [1][2][3] and document clustering [4][5][6][7][8], each of these research projects targeted only one type of document. As a result, the existing methods are unsuitable for topic extraction and clustering across different types of documents. Thus, topic extraction independent of document type and topic-centered management of documents are needed to achieve clever document management in accordance with how the information is to be used.

This paper describes algorithms that have been developed for topic extraction across document types and methods that support document management based on the topic.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 580–590, 2009.
© Springer-Verlag Berlin Heidelberg 2009

2 Problems in Document Management Related to Purpose of Use and Support Policies

2.1 Research Targets

The target of this research was intelligent activities in a network environment such as research activities, exploratory learning, and cooperative software development. In these activities, it is important to discover and efficiently utilize the desired documents in accordance with how the information is to be used. Since users accumulate various types of documents and in increasing quantities as activities progress, it is often difficult to identify the useful documents from the many documents accumulated.

The search for a desired document is enforced by focusing on the purposes of use. The topics hidden in the numerous documents are used here as indicators of the purpose of use. However, the elements of the documents differ with document type (e-mail, Web page, and so on). Consequently, topic extraction is harder when various types of documents are involved.

Moreover, the software applications differ with the document type; for example, mail client software is used for e-mail, and a Web browser is used for Web pages. Therefore, users can become bewildered as they come and go among various applications. Accordingly, it can be difficult for users to manage different types of documents while relating them sufficiently. This can make it difficult to grasp the progress of a project and the relationships among documents.

Consequently, support is needed for clever document management that enables users to locate and utilize useful documents regardless of the document type and application. This research targets document management by people working together such as on a laboratory research project. We assume that all of them use e-mail and use several types of documents.

2.2 Related Research

Many methods for topic extraction and document clustering have been reported. Hamasaki et al. described a method for discovering networks of common topics from bookmarks [1]. This method uses the hierarchical structure of bookmarks and identifies potentially useful pages by investigating the communities of topics among users. Sekiguchi et al. developed a method for extracting topics that uses the characteristics of utterances in weblogs [2]. Moreover, a trial in which networks of human relationships were extracted from information on the Web [9] has been reported.

The research on document clustering includes the work by Yanai et al. on automatic image clustering [4]. They developed a method for clustering natural images from the real world on the basis of learning from images gathered from the Web. Iyama et al. developed a system for clustering Web pages on the basis of their characteristics and for providing the results to users [5]. These clustering methods are unable to extract the topic across various types of documents and to cluster different types of documents since they each target only one type of document.

Systems that support document management and sharing have also been developed. Sano et al. described an information sharing system [10] that analyses the bookmarks of many users and uses the results to recommend a URI for users with similar interests. An interesting direction for research is the support of not only topic extraction but also of information sharing based on it. Moreover, Gmail [11], Google's e-mail program, is a strong tool that groups e-mails automatically on the basis of their reply relationships and supports management of e-mails on basis of the extracted groups. Nevertheless, most available systems support only one type of document. Therefore, managing different types of document efficiently is difficult. Furthermore, support for identifying the relationships among documents has not been investigated enough although such an identification would also contribute to understanding the work processes.

2.3 Issues and Support Policies

From the above discussion, we can identify three issues that need to be addressed in order to realize support for document management and utilization in accordance with how the information contained in them will be used.

1. Difficulty of topic extraction from different types of documents.
2. Difficulty of managing different types of documents while relating them.
3. Difficulty of understanding the relationships among documents.

To resolve these issues, algorithms have been developed for extracting the topic across a variety of document types. Moreover, an adaptive document management system has been developed that constructs its own functions and interfaces and changes them in accordance with the target types of documents and the purposes for which the information they contain will be used. This system has a function to assist understanding of work processes, document transitions, and their update circumstances by visualizing the relationships among documents related to a topic.

Here, topic extraction methods can be classified roughly into two types: 1) those that use a natural language based approach to either extract the contents of documents in a semantically constrained way or summarize them (e.g., [12]); 2) those that extract groups of documents that correspond to a topic (e.g., [13]). In this research, the latter approach was used as it aims to realize clever management and utilization of documents in accordance with how the information they contain will be used. The aim is to realize clever management by developing a topic-centered adaptive document management system (as described in Section 5).

3 Algorithm

3.1 Overview

A topic is represented by a group of documents corresponding to that topic. Each topic is considered to have various attributes (members, keywords, activity period, and so on). The set of these attributes, which expresses the feature of the topic, is called the "topic object." When a new document is acquired, the candidate topic to which it belongs is estimated on the basis of the coincidence rates, which are calculated by comparing the attributes of the new document with those of the existing topic objects. The coincidence rate is basically the general state of whether two attributes agree or not. The calculation methods differ for each combination of attributes, and they have been defined temporarily on the basis of an initial investigation.

Figure 1 illustrates the topic extraction algorithm. First, e-mails are grouped on the basis of the reply relationships. The resulting e-mail groups are used as a basic set of documents that expresses the topic. If there is more than one topic (group of documents), the attributes shown in Table 1 are extracted from each group. Each topic thereby acquires a set of elements that express its features, i.e., the topic object.

When a new document is acquired, the attributes specified beforehand in accordance with the types of documents (Table 1) are initially extracted. For the present time, e-mails, shared bookmarks, and PDF files are supported. If there is more than one topic, the relationship degree, which expresses the strength of the relationship between the new document and each topic, is calculated by comparing the attributes of both. If the calculated relationship degree exceeds a specified threshold, the topic with the maximum relationship degree is selected as the candidate topic for the new document. Conversely, if the relationship degrees for all topics fall below the threshold, the new document is judged to not belong to any topic.

Furthermore, if more than a specified number of e-mails belong to no topic before the calculation of the relationship degree, grouping of the e-mails is repeated. In this manner, new topics appearing after the initial grouping are handled.

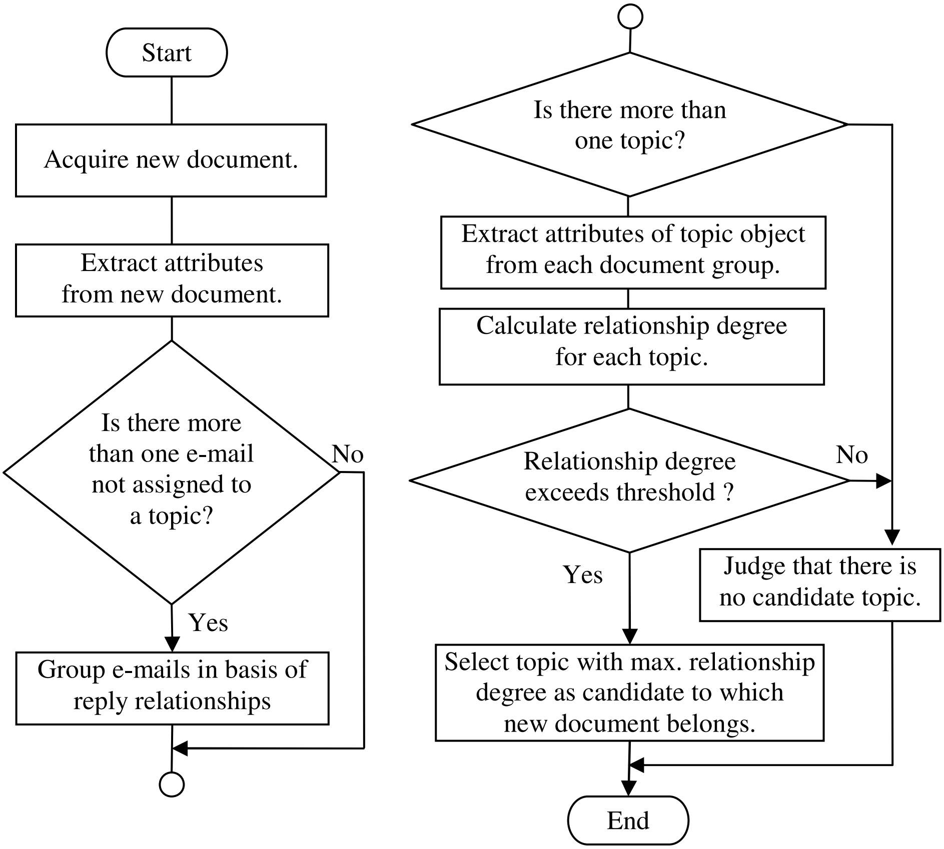
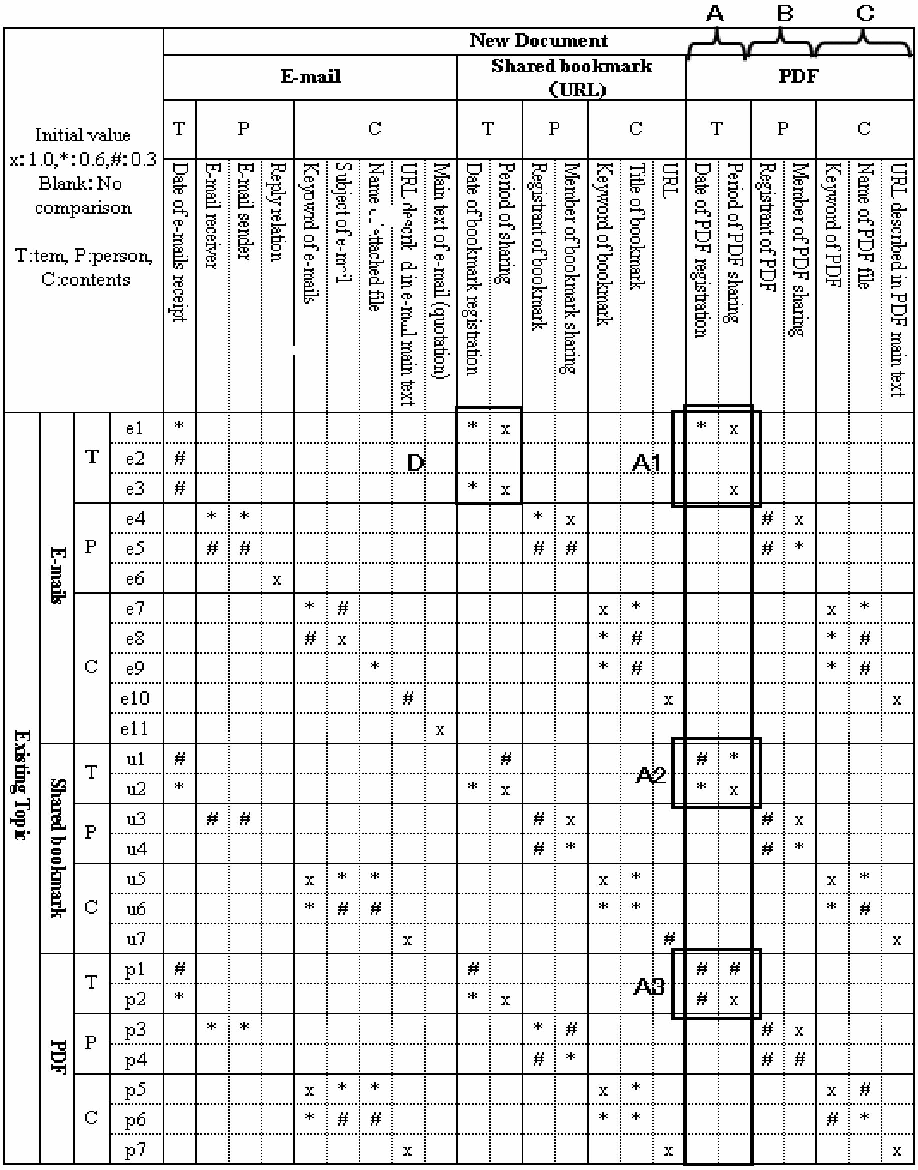

Start
Acquire new document.
Extract attributes from new document.
Is there more than one e-mail not assigned to a topic?
No
Yes
Group e-mails in basis of reply relationships
Is there more than one topic?
Extract attributes of topic object from each document group.
Calculate relationship degree for each topic.
Relationship degree exceeds threshold ?
No
Yes
Judge that there is no candidate topic.
Select topic with max. relationship degree as candidate to which new document belongs.
End

Fig. 1. Topic extraction algorithm

Table 1. Relationships among attributes of topic objects


A
B
C
D
A1
A2
A3

Initial value x:1.0,*:0.6,#:0.3 Blank: No comparison T:tem, P:person, C:contents				New Document																						
				E-mail									Shared bookmark (URL)							PDF						
				T	P			C					T		P		C			T		P		C		
				Date of e-mails receipt	E-mail receiver	E-mail sender	Reply relation	Keyword of e-mails	Subject of e-mail	Name of attached file	URL described in e-mail main text	Main text of e-mail (quotation)	Date of bookmark registration	Period of sharing	Registrant of bookmark	Member of bookmark sharing	Keyword of bookmark	Title of bookmark	URL	Date of PDF registration	Period of PDF sharing	Registrant of PDF	Member of PDF sharing	Keyword of PDF	Name of PDF file	URL described in PDF main text
Existing Topic	E-mails	T	e1	*									*	x						*	x					
			e2	#																						
			e3	#									*	x							x					
		P	e4		*	*									*	x						#	x			
			e5		#	#									#	#						#	*			
			e6				x																			
		C	e7					*	#								x	*						x	*	
			e8					#	x								*	#						*	#	
			e9							*							*	#						*	#	
			e10								#								x							x
			e11									x														
	Shared bookmark	T	u1	#										#						#	*					
			u2	*									*	x						*	x					
		P	u3		#	#									#	x						#	x			
			u4												#	*						#	*			
		C	u5					x	*	*							x	*						x	*	
			u6					*	#	#							*	*						*	#	
			u7								x								#							x
	PDF	T	p1	#									#							#	#					
			p2	*									*	x						#	x					
		P	p3		*	*									*	#						#	x			
			p4												#	*						#	#			
		C	p5					x	*	*							x	*						x	#	
			p6					*	#	#							*	*						#	*	
			p7								x								x							x

e1:Period of e-mails, **e2**:Average of e-mail reply, **e3**:Active period of e-mail, **e4**:Member of e-mail, **e5**:Key person, **e6**:Reply relationships, **e7**:Keyword of e-mail, **e8**:Subject of e-mail, **e9**:Name of attached file, **e10**:URLs in e-mail main text, **e11**:Main text of e-mail (quotation), **u1**:Bookmark registration date, **u2**: Bookmark sharing period, **u3**:Member of bookmark sharing, **u4**:Registrant of bookmark, **u5**:Keyword of bookmark, **u6**:Title of bookmark, **u7**:URL, **p1**:Date of PDF registration, **p2**:Period of PDF registration, **p3**:Member of PDF sharing, **p4**:Registrant of PDF, **p5**:Keyword of PDF, **p6**:Name of PDF file, **p7**:URL described in PDF main text.

3.2 E-mail Grouping

The focus here is on e-mail as this is the most common type of document and most everyone uses it. Groups of e-mails are used to express the basis of topics. Each e-mail has a unique ID called a Message-ID. Moreover, either “In-reply-to” or “Reference” information is added to the header of an e-mail sent in reply to a previous e-mail.

“In-reply-to” indicates the original e-mail by its Message-ID. “Reference” expresses a sequence of e-mails included in either previously sent e-mails or replied-to ones by their Message-IDs. Generally, at least one of these information fields is added into to the header of a reply e-mail. Here, e-mails are grouped by analyzing them.

3.3 Extraction of Topic Object

Several attributes are extracted from each group of e-mails created by analyzing the coincidence of the reply relationships. The extracted attributes are used as factors of the topic object as an initial step.

- **Topic period**: For each group, extract earliest and latest e-mail and use their creation dates as dates when the topic started and ended.
- **Average reply interval**: For each group, calculate average time from when e-mail was sent to when a reply to it was sent.
- **Keyword**: For each group, extract the top five characteristic words from the main texts of the e-mails.
- **Member**: For each group, extract the sets of senders and receivers.
- **Attached file**: For each group, extract the set of attached files.
- **Subject**: For each group, extract the set of subjects.
- **Reply Info.**: For each group, extract the “Message-ID” and either the “Reference” or “In-reply-to” from the e-mail headers.
- **URL**: For each group, extract the URLs from the main texts of the e-mails.

3.4 Calculation of Relationship Degree

This section describes the method used to calculate the relationship degree used for judging the topic to which a new document belongs.

When a new document is acquired, the specified attributes are extracted. These typical attributes were selected on the basis of previous investigation and experience. Then, the relationship degree is calculated by comparing the extracted attributes and the factors of each topic object. However, there are no comparable attributes between the new document and topic object when the document type of new document is not included in the target object. For instance, if a bookmark is registered immediately after initial grouping comparison is impossible since a topic object consists only of e-mails. This problem was solved by investigating the relationships among the factors of the various types of documents that went into building the topic object and then specifying the initial relationships, as shown in Table 1.

The main attributes that characterize the topics vary by topic and user. As a countermeasure against this problem, three meta-divisions were introduced for the topic object factors: term, person, and contents (Table 1).

The relationship degree of document *i* to document *j*, *rel(i,j)*, is calculated using equation (1). First, the coincidence rates between the combinations of factors are calculated. The target combinations and their initial weights are specified as shown in Table 1. The relationship degree is expressed as a value that is more than 0 and less than 1. The sum of the coincidence rates between the new document and the existing topics is finally calculated using normalized weights based on preliminary experiments for every area of the combinations of meta-divisions and the types of documents (e.g., when type of new document is PDF, A1 in Table 1).

Then, the calculated sum is multiplied by the ratio of the number of comparison targets to the number of all documents in the topic. The same calculation is done for areas A2 and A3 as well. The sum of these three values is the coincidence rate for each meta-division (e.g., A in Table 1): $agr_{ter}(i,j)$, $agr_{per}(i,j)$, $agr_{con}(i,j)$.

Finally, relationship degree *rel(i, j)* is obtained by summing the coincidence rates of the three meta-divisions multiplied with their respective weights: $w_{ter}(j)$, $w_{per}(j)$ $w_{con}(j)$ Although documents could belong to more than one topic, it is assumed that each document has only one candidate topic.

$$rel(i,j) = w_{ter}(j) \cdot agr_{ter}(i,j) + w_{per}(j) \cdot agr_{per}(i,j) + w_{con}(j) \cdot agr_{con}(i,j)$$

$$1 \geq w_{ter}(j), w_{per}(j), w_{con}(j) \geq 0,$$

$$w_{ter}(j) + w_{per}(j) + w_{con}(j) = 1 \tag{1}$$

4 Methods for Updating Relationships

4.1 Overview

Document classification differs by user and by circumstance. In one case, a user might cluster documents on the basis of their contents such as when documents related to a research theme are gathered. In another case, a user might arrange documents on the basis of the creator or receiver such as when e-mails received from and/or sent to a particular person are gathered. In a third case, a user might manage documents on the basis of time such as when documents created during a certain period are gathered.

An algorithm for topic extraction should be able to handle these various cases. The method described in this section can be used to update the relationships among the attributes of the topic object and meta-divisions of the attributes.

4.2 Method for Updating Relationships among Attributes of Topic Object

As mentioned above, relationships among attributes of topic objects differ with the topic and the user. Therefore, in the method described here, updating is done topic by topic.

Consider this example: an attribute is selected from the “term” factors for e-mail as the existing document and an attribute is selected from the “term” factors of bookmarks as the new document (D in Table 1). Initially, two types of attributes are selected. If the combination of the selected two attributes is marked (weighted) in Table 1, the relationship between those two attributes is evaluated in the same way as the relationship degree is calculated. Next, how much the attributes of the term divisions (Registration date, Sharing period) of all bookmarks belonging to the same topic coincide with those of e-mails (Period of e-mails, Active period) is evaluated. The average of that rate becomes the new relationship value. Specifically, the following coincidence rates are calculated, except for the blank cells in Table 1.

- Period of e-mails includes bookmark registration date or not (Includes: 1, Not included: 0).
- Rate of period of e-mails includes bookmark sharing period (number of overlapping days between period of e-mails and bookmark sharing period / total number of days for period of e-mails).
- A bookmark is registered during active period of e-mails (Registered: 1, Not registered: 0).
- Rate of active period of e-mails includes bookmark sharing period (number of overlapping days between active period of e-mails and bookmark sharing period / total number of days for active period of e-mails).

That is, evaluation in this updating examines the rate of influence that coincidence between each factor of topic object affects the belonging of documents to the topic, while calculation of relationship degree for a new document is used to judge the candidate topic to which the new document belongs.

For combination of other types of documents, the relationships among them are updated in the same way. The timing for the updating is determined on the basis of the date of the last updating, the number of accumulated documents, and so on.

4.3 Method for Updating Relationships among Meta-divisions

The updating of the relationships among meta-divisions of the attributes of topic objects involves term, person, and contents. This updating is done for each topic and user, the same in updating the relationships among attributes. Specifically, updating is done in accordance with the following procedure.

(1) Select one document from all documents belonging to the target topic.
(2) Calculate the coincidence between the attributes of the selected document and those of topic object same to calculation of relationship degree (more than 0 and less than 1).
(3) Calculate the average coincidences for the three meta-divisions (more than 0 and less than 1).
(4) Repeat steps (1) to (3) for all documents and then calculate the averages for the three meta-divisions.
(5) Replace the rate for the averages of the three meta-divisions with the new calculated averages.

This updating and replacement should improve the accuracy of topic extraction across different of types of documents and contribute to achieving topic-centered document management. The timing and frequency of the updating should be investigated since it requires calculation costs.

5 Topic-Centered Document Management System

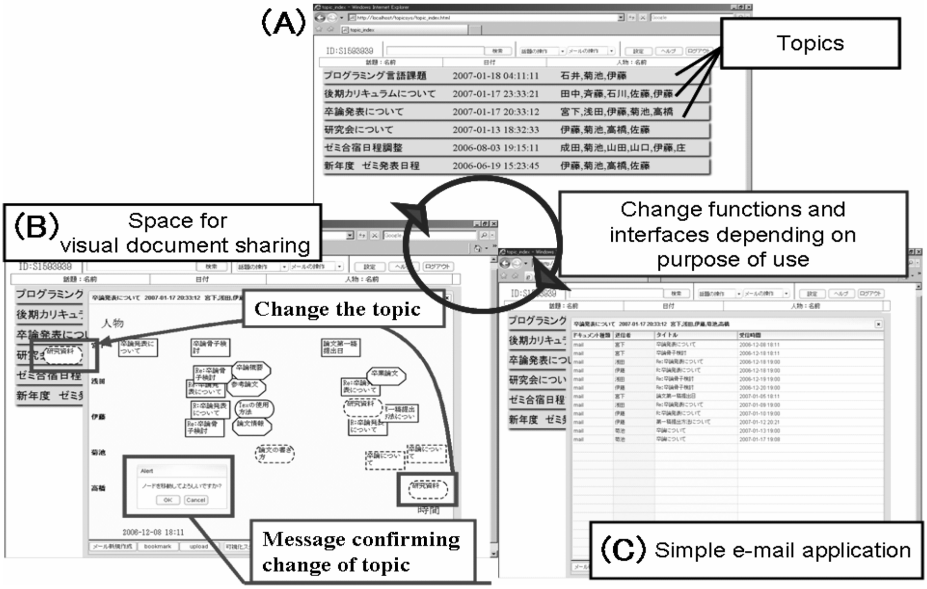
This research aims at developing a topic extraction method that can handle a mixture of different-types electronic documents and at developing an electronic document management system based on grouping by topics. This section describes the outline and basic idea of such a system.

In this system, the elements indicating the topic are automatically extracted from electronic documents like e-mails, and the extracted topics are used as a basic unit for various operations in document management. For example, at the time of system startup, as shown in Figure 2(A), rather than individual documents, topics are presented by listing characteristic words extracted from documents or elements of members.

The management area corresponding to the selected topic is shown, and then several types of documents belonging to the topic are displayed as a node with a different shape and a different color (Figure 2(B)). When a newly arrived document is judged to belong to a topic, it is shown as a special node in the management area of that topic. The user can change the topic judged by the system through interactive operation using a semi-automatic system.

Moreover, to help the user grasp the relationships among documents corresponding to various viewpoints, the visual presentation of relationships among documents is done considering various factors (e.g., time, creator, types of documents). This visualization can also contribute to grasp processes of works and to share them.

Furthermore, there is adaptive construction of functions and interfaces corresponding to the situation. For example, the system sometimes works as simple e-mail


(A)
Topics
(B) Space for visual document sharing
Change functions and interfaces depending on purpose of use
Change the topic
Message confirming change of topic
(C) Simple e-mail application
topic_index - Windows Internet Explorer
http://localhost/topicsys/topic_index.html
ID:S1593939
検索
話題の操作
メールの操作
設定
ヘルプ
ログアウト
人物
時間
Alert
ノードを移動してよろしいですか?
OK
Cancel
2006-12-08 18:11

話題：名前	日付	人物：名前
プログラミング言語課題	2007-01-18 04:11:11	石井,菊池,伊藤
後期カリキュラムについて	2007-01-17 23:33:21	田中,斉藤,石川,佐藤,伊藤
卒論発表について	2007-01-17 20:33:12	宮下,浅田,伊藤,菊池,高橋
研究会について	2007-01-13 18:32:33	伊藤,菊池,高橋,佐藤
ゼミ合宿日程調整	2006-08-03 19:15:11	成田,菊池,山田,山口,伊藤,庄
新年度　ゼミ発表日程	2006-06-19 15:23:45	伊藤,菊池,高橋,佐藤

卒論発表について 2007-01-17 20:33:12 宮下,浅田,伊藤,菊池,高橋

ドキュメント種類	送信者	タイトル	受信時間
mail	宮下	卒論発表について	2006-12-08 18:11
mail	宮下	卒論骨子検討	2006-12-18 18:11
mail	浅田	Re:卒論発表について	2006-12-18 19:00
mail	伊藤	R:卒論発表について	2006-12-18 19:00
mail	浅田	Re:卒論骨子検討	2006-12-19 19:00
mail	伊藤	Re:卒論骨子検討	2006-12-20 19:00
mail	宮下	論文第一稿提出日	2007-01-05 18:11
mail	浅田	Re:卒論発表について	2007-01-09 19:00
mail	伊藤	R:卒論発表について	2007-01-10 19:00
mail	伊藤	第一稿提出方法について	2007-01-12 20:21
mail	菊池	卒論について	2007-01-13 19:00
mail	菊池	卒論について	2007-01-17 19:08

Fig. 2. Topic-centered document management system

software (Figure 2(C)) and it other times evolves into an integrated shared environment in which documents of various types are treated (Figure 2(B)).

In conventional styles, users are often bewildered as they come and go among various applications because available applications depended on types of documents. The system described here enables new relationship between systems and a user so that systems adapt themselves to the needs of the user.

6 Conclusions

This paper describes algorithms for topic extraction across document types using topic object. In addition to details of the algorithm for judging topic to which new document belongs, methods for updating relationships among attributes of topic object and among meta-divisions of the attributes are also described. Finally, an adaptive system to realize topic-centered document management is proposed.

Future work includes designing and developing a prototype system and evaluating the topic extraction algorithm.

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan and by Japan Society for the Promotion of Science under Grant-in-Aid for Young Scientists (B) (No.20700632), Grant-in-Aid for Scientific Research (B) (No.21300304) (No. 19300272) respectively.

References

1. Hamasaki, M., Takeda, H., Matsuzuka, T., Taniguchi, Y., Kono, Y., Kidode, M.: A Method of Discovery of Shared Topic Networks among People from WWW Bookmarks and Its Evaluations. Trans. JSAI 17(3), 276–284 (2002) (in Japanese)
2. Sekiguchi, Y., Kawashima, H., Okuda, H., Oku, M.: Topic Detection from Blog Documents Using Bloggers' Interest. DBSJ Letters 5(1), 9–12 (2006)
3. Lamping, J., Rao, R., Pirolli, P.: A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: Proc. of the ACM Conference on Human Factors in Computing Systems, pp. 401–408. Addison-Wesley, Reading (1995)
4. Yanai, K.: Mining Visual Knowledge on the World Wide Web for G eneric Image Classification. Trans. JSAI 19(5), 429–439 (2004) (in Japanese)
5. Iyama, A., Sunayama, W., Yachida, M.: Topic Collection Support by Clustering Web Pages based on Topical Independence. Trans. JSAI 19(6), 561–570 (2004) (in Japanese)
6. Crawford, E., Kay, J., McCreath, E.: Automatic Induction of Rule for e-mail Classification. In: Proc. of the 6th Australasian Document Computing Symposium (2001)
7. Ueda, Y., Narita, H., Kato, N., Hayashi, K., Nambo, H., Kimura, H.: An Automatic Email Distribution by Using Text Mining and Reinforcement Learning. Trans. IEICE Inf. & Syst. J87-D1(10), 887–898 (2004) (in Japanese)
8. Balter, O., Sidner, C.L.: Bifrost Inbox Organizer: Giving users control over the inbox. In: Proc. of the Second Nordic Congerence on Human-Computer Interaction (2002)

9. Matsuo, Y., Tomobe, H., Hasida, K., Nakashima, H., Ishizuka, M.: Social Network Extraction from the Web information. Trans. JSAI 20(1), 46–56 (2005) (in Japanese)
10. Sano, K., Sayama, H.: BisNet:An Information Sharing System Using Bookmarks of Web Browsers. Trans. JSAI 20(4), 281–288 (2005)
11. Gmail, `http://mail.google.com/mail/`
12. Takaki, T., Fujii, A., Ishikawa, T.: Associative Document Retrieval by Query Subtopic Analysis and its Application to Patent Search. IPSJ Journal 46(4), 1074–1081 (2005)
13. Toyoda, M., Yoshida, S., Kitsuregawa, M.: Web Community Chart: A Tool for Navigating Numerous Web Pages by Related Topics. Trans. IEICE Inf. & Syst. J87-D1(2), 256–265 (2004)