

A Dynamic Grouping Technique for Distributing Codified-Knowledge in Large Organizations

J. Leon Zhao
Department of MIS
University of Arizona
Tucson, AZ 85721

Akhil Kumar
College of Business, Box 419
University of Colorado
Boulder, CO 80309-0419

Edward A. Stohr
Stern School of Business
New York University
New York, NY 10012, USA

Abstract

Information overload is a major problem plaguing the “mailing lists” approach to knowledge distribution. A recent trend towards resolving the problem is to distribute information to organizational members according to their individual needs. In this paper, we propose a new technique, referred to as “dynamic grouping”, for distributing codified knowledge in large organizations. To enable dynamic grouping, we develop a data structure called “organizational concept space” (OCS) consisting of a *similarity network* and an *interest matrix*. One major advantage of the dynamic grouping technique is that it can reduce information overload while avoiding information starvation by accommodating different levels of user demand for knowledge.

1. Introduction

Distributing codified-knowledge to the right people at the right time is a fundamental component of knowledge management in modern organizations (Zack, 1999). By codified-knowledge, we refer to the body of knowledge that can be communicated in digital form and is important to the coordination of corporate activities. However, the current state of the art in information distribution technologies still lags behind corporate needs (Foltz and Dumais, 1992; Goldberg et. al, 19992). The most prevalent mechanism of information distribution is to send messages via mailing lists. However, e-mail distribution suffers from the twin problems of information overload and information starvation (Hall, 1998; Zhao, Kumar, and Stohr, 2000).

New approaches to organizational information distribution have been developed recently based on profiling user interests and distributing messages to interested users only (Kindo et al., 1997). Information filters are developed to present each user with all information relevant to that user’s information needs. User profiles are needed to model users’ interests in order to determine what information the user might need. However,

building user models is difficult because it is not easy for users to specify what those interests are (Stadnyk and Kass, 1992; Oard, 1996). One of the major obstacles is the vocabulary problem (Chen et. al, 1996). The terms contained in the information available are different from those the user would use to specify his or her interests. Besides the vocabulary problem, other issues related to the modeling of user interests include goals (how information relates to user interests), message types (message classes such as call for papers, book reviews, conference announcements), message characteristics (length of message, author of the message), and so on.

Another approach relevant to this study is *conceptual clustering*, in which documents are classified based on the terms they contain and queries are processed based on the terms specified by the user (Munoz, 1997). In conceptual clustering, documents are specified as vectors in the multidimensional space of keywords and are clustered according to the frequency of keywords appearing in them. The main advantage of the vector space model is that the development of conceptual clusters is automated using computational procedures. Similarly, a more comprehensive approach, termed *concept space*, has been proposed to create meaningful and understandable domain-specific networks of terms and weighted associations, automatically computed from thousands of documents (Chen et al., 1996).

In this study, we extend conceptual clustering techniques by incorporating organizational information directly in the data structure to create an *organizational concept space*. We then develop algorithms that can dynamically group users relative to the information to be distributed. The resulting technique is called *the dynamic grouping technique* for distributing codified knowledge.

2. Message Distribution Based on Dynamic Grouping of Users

A major drawback of the mailing lists approach, used extensively in corporations, is the “uniformity

assumption” (Zhao, Kumar, and Stohr, 2000), which supposes that all users of a mailing list have the same knowledge needs and should therefore receive the same messages. By contrast, in dynamic grouping, users are grouped with respect to each particular message and the grouping changes from message to message. It is in this sense that we refer to our technique as dynamic grouping as illustrated in Figure 1.

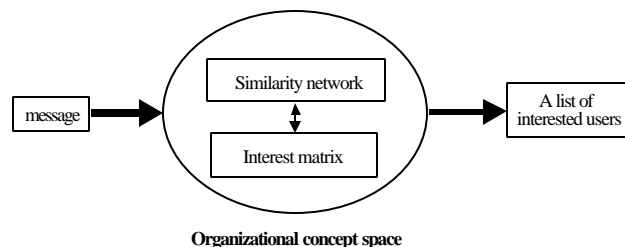


Figure 1. The dynamic grouping technique.

In this approach, a message is fed through the organizational concept space, and a list of users who are interested in the message is derived. The key data structure of the approach is the *organizational concept space*, which consists of a similarity network and an interest matrix. A *similarity network* is a collection of similarity sets that form a network based on levels of generalization/specialization. Figure 2 illustrates a simple similarity network.

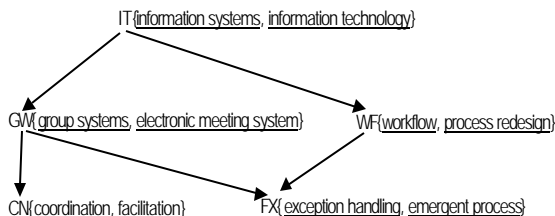


Figure 2. A simple similarity network.

A *similarity set* is a collection of concepts that are closely related semantically. For instance, “group systems” and “electronic meeting systems” are two concepts that are used interchangeably by many people. The concepts can be associated to the similarity set with a membership value between 0 and 1 denoting the degree of association (not shown in Figure 2 for simplicity).

An *interest matrix* is a 2-dimensional matrix with “concept” as one dimension and “user” as the other (see Figure 3). The entries of the matrix indicate the level of interest of the user in the concept. These are also referred to as *interest*

indicators. An interest indicator is a continuous variable from 0 to 1, which indicates the *degree of interest* of the user in the concept.

To determine the users interested in a message, the message contents are described by a number of representative concepts that are either provided by the message sender, or derived from the organizational concept space. The *message representative concepts* are then matched to the organizational concept space to determine the users potentially interested in the message. Next, we delineate the detailed data structures and algorithms for deriving the group of users for a given message.

3. Organizational Concept Space

An *organizational concept space (OCS)* stores the user interests in an organization via an interest matrix and a similarity network. The user interest matrix contains the correspondence between all concepts and users in an organization, and the similarity network organizes all concepts (or topics) into a network of similarity sets, each of which contains a set of similar concepts.

	T_1	T_2	T_3	...	T_{n-2}	T_{n-1}	T_n
U_1	0.4	0	0.6	...	0.7	0.2	0
U_2	0.3	0	0	...	0.5	0	0
U_3	0	0.2	0	...	0	0.8	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
U_n	0.2	0	0	...	0.9	0	0

M : The message
 T_j : The j^{th} topic in the OCS
 U_i : The i^{th} user
 $\{U_i, T_j\}$: Level of interest of user i in topic j
 T_j : The j^{th} topic in message M
 U_i : A user interested in message M

Figure 3. An interest matrix.

Figure 3 illustrates an interest matrix where each row corresponds to a user U_i and each column corresponds to a concept (or topic) T_j . The intersection of a row and a column indicates the extent of a user’s interest in the concept (between 0 and 1). As shown in the figure, the concepts of a message M are matched to the concepts of the interest matrix, and then are projected to the users that are interested in the concepts of the message where a nonzero indicator is found. The matched topics and user are marked with circles in Figure 3.

Although an interest matrix is effective for determining the users interested in the concepts of a message, it does not provide a means to deal with the vocabulary problem (Chen et. al, 1996). This

consists of a number of subproblems such as synonyms, spelling variations, and semantic subsumption. To this end, we propose a technique referred to as a “similarity network” as shown in Figure 4.

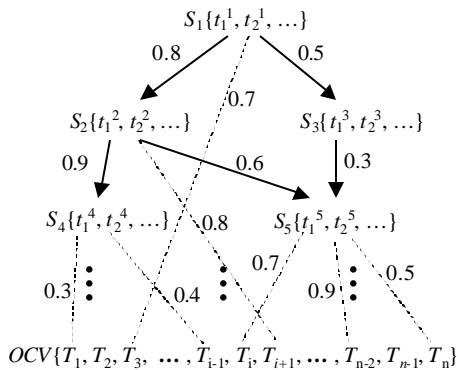


Figure 4. Similarity network and concept extension.

A similarity network is a directed network comprised of two types of nodes, body nodes and leaf nodes. A body node is a similarity set containing a collection of concepts that are similar in semantics, and a leaf node is a concept (or topic). Each leaf node is linked to one or more body nodes via a dashed line, and each body node is linked to one or more parent body nodes via a thick arrow, resulting in a network. All leaf nodes are organized into a list called the organizational concept vector (OCV). The decimal values on the arrows indicate the strength of the associations between two body nodes and the degree of membership of a leaf node in its similarity set, respectively.

The similarity network can be used to determine concepts matching those present in message M by “extending” the matching sets in a novel way. This process, referred to as *concept extension*, is as follows:

- *Direct match*: Given a set of concepts representing a message, say $M(p_1, p_2, \dots, p_m)$, the corresponding concepts (or topics T_j s) in the OCV are called matching concepts if there is a direct match between M and OCV.
- *Level zero indirect match*: Then, each matching concept from the direct match step is linked to its similarity set following the link between itself and its parent set. All concepts in its parent set are considered to be matching concepts due to an indirect match. Because the new concepts are from the parent set, they are

referred to as level zero *indirect* matching concepts.

- *Higher level indirect match*: The similarity network can be used to propagate the matching process along paths of the network. That is, a path can be found through the network from a concept present in the message to another concept of interest to the user. To distinguish between different levels of indirect match, we refer to a higher level match as “plus level X ” indirect match and a lower level match as “minus level Y ” indirect match, in terms of the relative distance of the matching set from the parent set of the original concept.

These matching schemes in the similarity network offer a meaningful way of assessing the strength of match between the user interest and the message. For example, a concept at a higher node of the network is more general than one at a lower node. Two concepts close to each other in the network will add more weight to the user interest than two that are far apart.

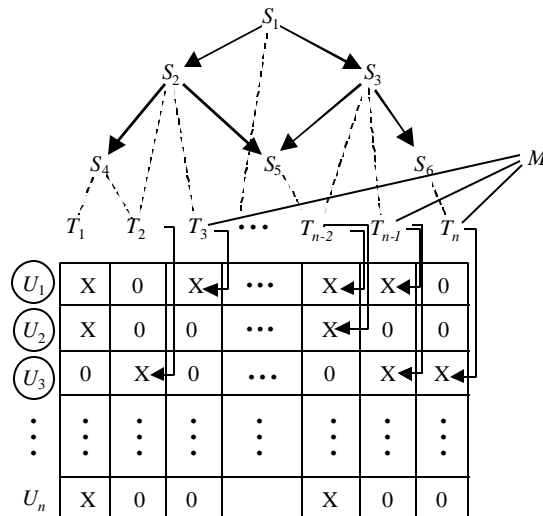


Figure 5. Extended matching via interest matrix and similarity network.

Figure 5 illustrates the integrated use of the interest matrix and the similarity network to determine the users interested in a particular message M . The X marks in Figure 4 are used to represent nonzero interest indicators.

A three-step procedure can be devised to conduct the matching process with respect to the similarity network:

1. *Direct match*: Given a message and its concepts, determine the direct matching concepts in the organizational concept vector, their parent similarity sets, and the users who are interested in the direct matching concepts. In Figure 5, we can see that user U_1 is interested directly in message M via concepts T_3 and T_{n-1} , and similarly, user U_3 is interested directly in M via concepts T_{n-1} and T_n .
2. *Level zero match*: Determine the concepts of the level zero match, i.e., the siblings of the concepts with a direct match that have the same parent sets. In Figure 5, user U_2 is interested in the message indirectly at level zero because of his/her interest in concept T_{n-2} , which is a sibling of T_{n-1} .
3. *Higher level matches*: Determine the concepts at higher levels of match, i.e., those contained in the similarity sets linked to the parent set.

Based on the results of the above matching process, the users who should be interested in the message can be derived. As mentioned above, users differ in their aggressiveness towards obtaining knowledge. Therefore, the users who are interested in each each message should be selected by applying different standards. The matching paradigm, i.e., direct vs. indirect matching plus the various levels of indirect matching outlined above provides a natural way to predict the level of user interest in the same message for various users.

For instance, conservative users might be selected only if there are a sufficient number of direct match concepts, moderate users if there are sufficient level zero matches of concepts, and aggressive users if there are additional concepts matched at higher levels of indirect match. These points of sufficiency can be determined by applying threshold values to the match ratio.

4. A Relational Implementation

4.1. Data Structure

To experiment with the proposed approach, we use a relational database system to implement a prototype. The advantages of relational prototyping are portability and scalability. While this may sacrifice some computational efficiency, a relational prototype is sufficient for our purpose of validating the proposed methodology. The relational schemas for the data structures are:

- Interest Matrix:

$IM(\underline{\text{User}}, \underline{\text{Topic}}, \text{Interest})$ where User contains the user ID, Topic contains the topic names, and Interest contains the level of interest, which can be either a binary or continuous variable depending on whether exact or inexact membership is used. The User and Topic are the composite primary key of the relation and are underlined.

- Demand Level:

$DL(\underline{\text{User}}, \text{Aggressiveness}, \text{Threshold})$ where User is the user ID, Aggressiveness contains the level of user demand for knowledge, and Threshold is an integer specifying the minimal number of matching topics for the overall interest in a message in the case of a simple interest matrix or a fraction specifying the minimal value for the overall interest in a message in the case of a fuzzy interest matrix. The values of Aggressiveness are “aggressive”, “moderate”, and “conservative”. The aggressiveness values are used to characterize different user demand levels when deciding the users to whom the message should be sent.

- Similarity Sets:

$SS(\underline{\text{Set}}, \underline{\text{Topic}}, \text{Membership})$ where Set is the set ID of a similarity set, Topic is the topic name in question, and Membership is a value from 0 to 1.

- Similarity Network:

$SN(\underline{\text{Parent}}, \underline{\text{Child}}, \text{Association})$ where Parent contains the set ID of a parent similarity set, Child contains the set ID of a child similarity set, and Association is a value from 0 to 1. Each tuple in the relation stores a network link and can be used to propagate the concepts by tracing these network links.

4.2. Computational Algorithms

Given the data structures in a relational model, the algorithm for determining user interests in a message can be specified using SQL statements as follows, assuming the topics (or concepts) for a specific message M has been given in a set *Topics*:

1. Find relevant similarity sets:

```
SELECT DISTINCT Set
FROM SS
```

```

WHERE Topic IN Topics
INTO Parent_Sets

SELECT DISTINCT Parent
FROM SN
WHERE Child IN Parent_Sets.Set
INTO Grandparent_Sets

SELECT DISTINCT Child
FROM SN
WHERE Parent IN Parent_Sets.Set
INTO Uncle_Sets

```

2. Find extended topics:

```

SELECT DISTINCT Topic
FROM SS
WHERE Set IN Parent_Sets.Set
INTO Sibling_topics

SELECT DISTINCT Topic
FROM SS
WHERE Set IN Grandparent_Sets
INTO Uncle_topics

SELECT DISTINCT Topic
FROM SS
WHERE Set IN Sibling_Sets
INTO Nephew_topics

```

Let the set *Conservative_topics* be *Topics*, the set *Moderate_topics* be the union of *Topics* and *Sibling_topics*, and the set *Aggressive_topics* be the union of *Moderate_topics*, *Uncle_topics* and *Nephew_topics*. These three sets of topics will be used for the three types of users, namely aggressive, moderate, and conservative users.

3. Find interested users:

```

SELECT User
FROM IM, DL
WHERE Topic IN Aggressive_topics
AND Aggressiveness = "Aggressive"
AND IM.User = DL.User
HAVING Interest(User) >= Threshold
GROUP BY User
UNION
SELECT User
FROM IM, DL
WHERE Topic IN Moderate_topics
AND Aggressiveness = "Moderate"
AND IM.User = DL.User
HAVING Interest(User) >= Threshold
GROUP BY User
UNION
SELECT User
FROM IM, DL
WHERE Topic IN Conservative_topics

```

```

AND Aggressiveness = "Conservative"
AND IM.User = DL.User
HAVING Interest(User) >= Threshold
GROUP BY User

```

Interest() in the SQL statements above is a function for computing the aggregate interest level of the user in the message using the following three estimation principles that summarize a similar approach for query processing in conceptual networks (Lucarella and Morara, 1991; Chen and Hornig, 1999):

- *The Sequence Principle*: If there is a chain from a topic t contained in the message to the user, the interest of the user in topic t should be the minimum of all weights along the chain.
- *The Parallelism Principle*: If there are two chains from a topic t contained in the message to the user, the interest of the user in topic t should be the maximum of the interest values derived from the two chains.
- *The Aggregation Principle*: If there are multiple topics contained in the message, the interest of the user in the message should be the maximum of the interest values in all topics.

Note that the above principles are based on the MIN-MAX algorithm of fuzzy logic (Chen and Hornig, 1999). To illustrate the computational principles, consider the following example in Figure 6. Given user U_i and message M . M contains two concepts t_1 and t_2 with membership 0.9 and 0.7 in its similarity set S_1 , respectively. Concepts t_1 and t_2 also have a "nephew" concept t_3 with a membership 0.8 in the child set of S_1 plus an "uncle" concept t_4 with a membership 0.7 in the parent set of S_1 . The user has interest indicators of 0.4, 0.7, 0.8 and 0.6 in $t_1, t_2, t_3,$ and $t_4,$ respectively.

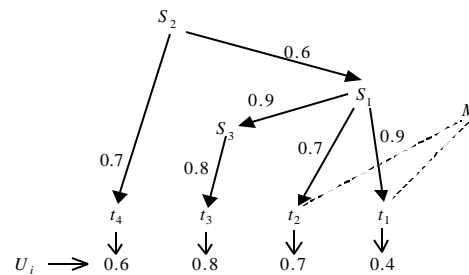


Figure 6. An interest aggregation example.

The overall interest of user U_i in message M is derived as follows using *the above* principles:

- (1) The user interest in t_1 directly = 0.4
- (2) The user interest in t_1 via the chain of t_3, S_3, S_1 ,
 $t_1 = \min(0.9, 0.9, 0.8, 0.8) = 0.8$ {Sequence}
- (3) The user interest in t_1 via the chain of t_4, S_2, S_1 ,
 $t_1 = \min(0.9, 0.6, 0.7, 0.6) = 0.6$
- (4) Therefore, the user interest in t_1 via all chains
 $= \max(0.4, 0.8, 0.6) = 0.8$ {Parallelism}
- (5) The user interest in t_2 directly = 0.7
- (6) The user interest in t_2 via the chain of t_3, S_3, S_1 ,
 $t_2 = \min(0.9, 0.9, 0.8, 0.8) = 0.8$
- (7) The user interest in t_2 via the chain of t_4, S_2, S_1 ,
 $t_2 = \min(0.9, 0.6, 0.7, 0.6) = 0.6$
- (8) Therefore, the user interest in t_2 via all chains
 $= \max(0.7, 0.8, 0.6) = 0.8$
- (9) Consequently, the user interest in message
 $M = \max(0.7, 0.8) = 0.8$ {Aggregation}

The interest level (0.8) will be compared with the threshold value of all the users to determine whether the user should be selected for retrieval of the message.

5. Conclusions

This extended abstract proposed a dynamic grouping approach for distributing codified knowledge in large organizations. Our goal is to reduce information overload while considering the different levels of user needs for information. We developed an organizational concept space technique that consists of a similarity network and an interest matrix. A relational model for the organizational concept space and the associated computational procedures are presented using SQL statements.

While we are optimistic about the usefulness of the proposed technique, many research issues remain unresolved. For instance, how to maintain the user interests is a major one. Although many institutions such as a university maintain researchers' interests and store working papers that contain keywords, these sources of user interests tend to be static and incomplete. The overhead of creating a similarity network can be considerably high. Therefore, some level of automation should be helpful.

We are currently working to validate the proposed dynamic grouping technique. To do so, we will develop metrics for evaluating the benefits of our proposed technique in comparison with conventional mailing lists. Examples of such metrics are the number of uninteresting messages

per user in a given time period, and the number of interesting messages missed per user in the same time period.

References

1. Chen, H. et al., "A parallel computing approach to creating engineering concept spaces for semantic retrieval: the Illinois digital library initiative project", *IEEE Trans on PAMI*, Vol. 18, No. 8, August 1996. Pp. 771-782.
2. Chen, S-M. and Horng, Y.-J., "Fuzzy query processing for document retrieval based on extended fuzzy concept networks", *IEEE Trans on SMC – Part B*, Vol. 29, No. 1, Feb. 1999.
3. Chesnais, P.R.; Mucklo, M.J.; Sheena, J.A. "The Fishwrap personalized news system", *2nd Intl Workshop on Integrated Multimedia Services to the Home*, 1995, Page(s): 275 -282
4. Foltz, P.W. and Dumais, S.T. "Personalized information delivery: an analysis of information filtering methods." *CACM*, 1992.
5. Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. "Using collaborative filtering to weave an information tapestry." *CACM*, 1992.
6. Hall, R.J. "How to avoid unwanted e-mail." *CACM*, March 1998, vol.41, (no.3):88-95.
7. Kindo, T.; Yoshida, H.; Morimoto, T.; Watanabe, T. "Adaptive personal information filtering system that organizes personal profiles automatically." *IJCAI 97*, Nagoya, Japan, 23-29 Aug. 1997, p. 716-21 vol.1.
8. Lucarella, D. and Morara, R. "FIRST: Fuzzy information retrieval system," *Journal of Information Science*, Vol. 17, pp. 81-91, 1991.
9. Munoz, A. "Compound key word generation from document databases using a hierarchical clustering ART model", *Intelligent Data Analysis*, Vol. 1, Number 1, January 1997.
10. Oard, D.W., "A Conceptual framework for text filtering," *U. Maryland*, Technical Report, CS-TR-3643, May, 1996.
11. Stadnyk, I. and Kass, R., "Modeling users' interests in information filters", *CACM*, 1992.
12. Zack, M. H., "Managing codified knowledge", *Sloan Mgmt Review*, Summer 1999, pp. 45-58.
13. Zhao, J. L., A. Kumar, and E. A. Stohr, "Workflow-centric Information Distribution through Email", *JMIS*, 2000 (forthcoming).