2008

Examining Granular Computing from a Modeling Perspective

Ying Xie
Kennesaw State University, yxie2@kennesaw.edu

Jayasimha R. Katukuri
EBay Inc

Vijay V. Raghavan
University of Louisiana at Lafayette

Examining Granular Computing from a Modeling Perspective

Ying Xie
Department of Computer Science and Information Systems
Kennesaw State University
1000 Chastain RD., Kennesaw, GA 30144
yxie2@kennesaw.edu

Jayasimha Katukuri, Vijay V. Raghavan
The Center for Advance Computer Studies
University of Louisiana at Lafayette
104 University Circle, Lafayette LA 70504
{jrk8907, raghavan}@cacs.louisiana.edu

Tom Johnsten
School of Computer and Information Sciences
University of South Alabama
Mobile, Alabama 36688
tjohnsten@usouthal.edu

***Abstract* - In this paper, we use a set of unified components to conduct granular modeling for problem solving paradigms in several fields of computing. Each identified component may represent a potential research direction in the field of granular computing. A granular computing model for information analysis is proposed. The model may suggest that granular computing is an instrument for implementing perception based computing based on numeric computing. In addition, a novel granular language modeling technique is proposed for information extraction from web pages. This paper also suggests that the study of data mining in the framework of granular computing may address the issues of interpretability and usage of discovered patterns.**

I. INTRODUCTION

Pioneer researchers in granular computing have studied this new emerging field from both high level philosophical or methodological perspectives [2,3,9,10] and from a mathematical instrumentation point of view [2,3,6,7,8,9,10,11,12,16]. It has been proposed that granular computing is an essential mechanism to achieve perception-based computing [2,3,9]. Some theories and instruments, such as fuzzy set, rough set, neighbourhood systems, and various logic languages and inference mechanisms are used to model information granulation and computing with granules [2,3,6,7,8,9,10,11,12,16]. In addition, granular computing has been used to categorize certain data mining tasks [6,13,14]. Nevertheless, there still lacks a relatively complete and detailed model of the granular computing paradigm that can be applied to characterize the problem solving process in different fields of computing. Therefore, this paper examines granular computing as a unified modeling framework to describe the problem solving paradigm in several different computing fields where human perceptions and understanding are essential. One of the advantages of modeling different domains from a unified granular perspective is that it enables us to extract crucial components of granular computing using a single formalism. Another advantage is that it helps to find previously unknown solutions for one domain based upon existing methods in another domain.

We first present a granular model for object-oriented software design. Second, we use the extracted components to present a relatively comprehensive granular model for information analysis. The developed model is then applied to web mining and computing with words. The model integrates perception based computing with numeric computing. This result suggests that granular computing may be an instrument for implementing perception based computing based on numeric computing.

II. GRANULAR MODELING OF OBJECT-ORIENTED SOFTWARE DESIGN

The objective of the object-oriented paradigm is to model an application domain based on how a human perceives the domain. This paradigm can be viewed as an instance of granular modeling. We can derive from the object-oriented paradigm some basic elements for granular modeling that can be shared by other domains. We start with an object, the basic unit of the object-oriented paradigm. An object can be viewed as a type of granule. It is described by its data members, member functions, and public interface. Therefore, a granule may be described by the current values of internal data (i.e., the status of the object), the operations applied to the internal data, and its relationships with other granules. We will refer to this type of representation of a granule as the ***basic representation***. A group of functionally interacting objects can form a software module. A software module is also a type of granule. Such granules have two representations. One representation is the basic

978-1-4244-2352-1/08/$25.00 ©2008 IEEE

representation that consists of integrated data values (status of the granule), integrated operations, and integrated interfaces. The other representation, referred to as *composite representation*, consists of objects, and relationships among objects. Furthermore, a group of functionally related modules form a subsystem, which is a granule at a coarser level; and, a group of subsystems form a system, which is a granule at an even coarser level. Hence, a hierarchical structure exists among granules at different granularity levels. Theoretically, all granules in a hierarchy structure have both a composite representation and a basic representation. However, for some granules at the upper level of a hierarchy, it may not be practical or necessary to specify a basic representation. A granule may have a ***hybrid representation,*** a combination of basic and composite representations.

For a granule, the conversion from a basic representation to a hybrid representation, then to a composite representation can be considered an instance of ***information abstraction***; while the conversion from the composite representation to a hybrid presentation, then to a basic representation can be an instance of ***information extraction***. The operation to form a granule from a set of granules defined at a lower level abstraction is referred to as ***granulation***. In contrast, the operation of decomposing a granule into a set of smaller granules is referred to as ***decomposition***. The process in which a group of objects interact with each other to perform a specific task is referred to as ***computing with granules***.

Object-oriented software design, therefore, can be described in terms of the following granular computing tasks:

- Specifying the composite representation of the largest granule – the system
- Conducting decomposition of the system into a group of smaller granules – subsystems
- Specifying the composite representation of a subsystem
- Conducting decomposition of the sub-system into a group of smaller granules – modules
- Specifying the composite representation of a module
- Conducting decomposition of the module into a group of even smaller granules – objects
- Specifying the basic representation of an object

Hence, some basic elements of granular computing can be extracted from the domain of object-oriented software design as listed below.

- Basic representation, composite representation, and hybrid representation of a granule
- Granulation and decomposition
- Information abstraction and information extraction
- Computing with granules
- Hierarchical structures of granules

III. GRANULAR MODELING OF INFORMATION ANALYSIS

Information processed by a human can be modelled as granules and the relationships among granules. We assume that an individual perceives a granule by different senses, such as size, shape, color, sound, and so on. Each sense generates a signal with a certain level of intensity in an individual's brain. Although the individual may not be able to explicitly quantify the level of intensity of a certain sense, internally the intensity level can be represented by a number. Therefore, the basic representation of a granule we discuss here is in certain form of numbers (The numbers here exclude the ones representing certain types of membership). For instance, a basic representation of a granule representing a time series includes a sequence of numbers. Please note that a granule may have multiple basic representations, each of which contains a different amount of information. Furthermore, we assume that the level of intensity of certain sense(s) perceived from a granule is immediately categorized into another granule by the brain. For instance, the level of intensity of sense of size may be categorized to a granule labelled as "big". Therefore, a granule has composite representation that describes the fact that this granule is composed of a group of other heterogeneous or homogeneous granules with relationships among those granules. A hybrid representation of a granule contains partially a group of constituent granules and partially numbers.

Based on the concepts of granules mentioned above, a typical information analysis process can be described in figure 1. The process starts with a repository of raw data. Then, by *granular modelling*, an overall granule (referred to as A in figure 1) that represents the data set is generated. Through the *decomposing* process, the overall granule is gradually decomposed into groups of granules (referred to as B in figure 1) at proper granularity in composite representation. Subsequently, granules in composite representation are converted into ones in basic representation (referred to as C in figure 1) by the *information extraction* process. Now, the granules in basic representation can be transformed into other types of basic representation (referred to as D in figure 1) by all kinds of numerical computing methods that can be applied to the types of numbers embedded in granules. In other words, these numerical methods update the status of one or more granules. Example transformation for a granule representing a time series may include filtering, sampling, wavelet transformation, matrix transformation, and so on.

For simplicity, the algorithms that generate granules at a higher level, such as clustering, classification and pattern extraction are excluded from this category; instead, they are included in the process of *information granulation*. Information granulation generates different types of clusters on numerical data based on their similarity, proximity, correlation, association, functional interactions or trends. The resulting clusters form groups of new granules having a basic representation (referred to as E in figure 1). These new granules are comprised of granules in D. The granules in D can be represented by the granules in E through the process of *information abstraction*. Information abstraction can be achieved through the processes of discretization and fuzzification of numerical values. The result of information abstraction is groups of granules having a composite or hybrid representation. Afterwards, a process, referred to as *computing with granules*, generates relationships among the granules in F and forms a granule-relationship map (referred to as G in figure 1). Finally the *semantic enrichment* process converts the granule-relationship map to a concept-relationship map that is comprehensible by humans. As one may notice that in figure 1, the groups of granules in composite representation referred to in B can be directly input to computing with granules model if they are already in suitable forms.

numerical computing. Therefore, building computational models and algorithms for these three components are the foundations of granular computing. Variety of data mining algorithms for clustering, classification, association mining, regression, and motif discovery can be used for *granulation*. From another perspective, if we view data mining as the granulation component in the whole picture of granular computing, it may help us to find formal ways to solve some challenging problems that data mining faces, such as interpretability of discovered patterns. The existing computing models that can be used for *computing with granules* include set theory, fuzzy logic, rough logic, probabilistic logic, possibility logic and others. More computing models are needed to deal with computing with granules that have complex relationship among their constituent granules. In the following sections, we will use the proposed granular model

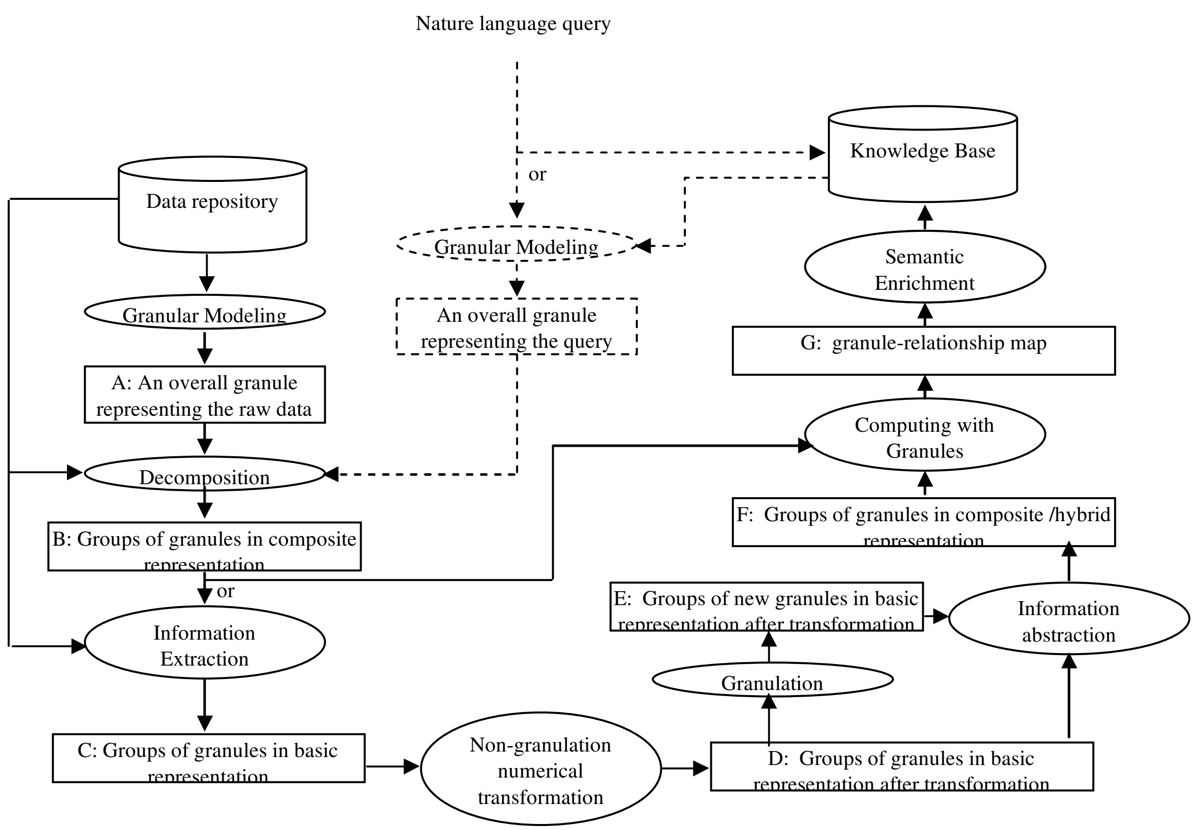

Figure 1 Granular Modeling for Information Analysis

As illustrated in figure 1, the major research areas of granular computing include: granular modeling, decomposition, information extraction, granulation, information extraction, computing with granules, and semantic enrichment. Among them, information extraction, granulation, and information abstraction are the interfaces of granular computing with for information analysis to categorize web mining and computing with words.

IV. GRANULAR MODELING OF WEB MINING

For Web Mining, the data repository is a collection of web pages, where each page contains a set of links. If we model the collection from a granular point of view, the overall granule can be represented as shown in figure 2. This granule can be decomposed into a group of smaller granules $\{G_{d1}, G_{d2}, \ldots G_{di}, \ldots\}$, where G_{di} contains page *di* and all the pages that link to *di*, as shown in figure 3.

Now, we can conduct information extraction to convert G_{di} from a composite representation to a basic representation by using the following probability model:

$$P(w \mid G_{di}) = P(w \mid di)P(di \mid G_{di}) + \sum_{j} P(w \mid pj)P(pj \mid Gdi)$$

where *j* is the *jth* parent page that links to *di*.

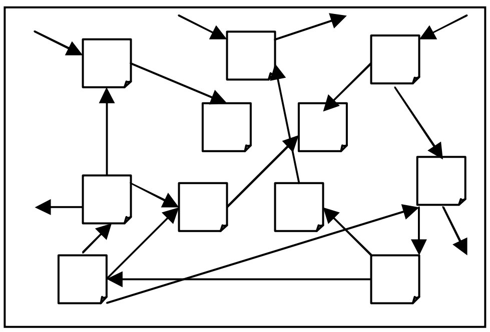

Figure 2 An overall granule modeling the whole web collection

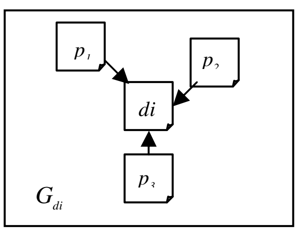

Figure 3 G_{Di} in composite representation

Basically, G_{di} is a granular version of page *di*. In other words, when analyzing page *di*, this granular version allows us to consider not only the page *di* itself, but also the parent pages that links to *di*. Actually *P(w|di)* is the formula for the language modeling [1] technique that estimates the probability of sampling a word *w* from the language model that is assumed to generate the page *di*. Here we propose the *Granular Language Model* $P(w| G_{di})$. This granular language model seamlessly integrates the property that the text surrounding the link (such as anchor text) from a parent page to a child page typically has an important indication of the content of the child page. The formula $P(w| p_j)$ represents the probability of sampling a word *w* from the text surrounding the link from *pj* to *di*. It can be estimated as: $P(w \mid p_j) = P(w \mid st)\dfrac{size(st)}{size(p_i)}$, where *st* represents the surrounding text of the links from p_j to d_i; *size(st)* represents the number of words in *st*, $size(p_j)$ represents the number of words in the page p_j.

The granular approach for information extraction can be easily extended to other domains where entities are part of a social network. This model reflects the fact that in order to evaluate an entity for a certain feature, one not only refers to the target entity itself, but also to the entities that have a relationship with the targeted entity.

The information extraction process converts a granule in a composite representation to a one in a basic representation. After that, various numerical transformation methods can be applied to the granules having a basic representation, such as latent semantic indexing (LSI) [5], generalized vector space modeling (GVSM) transformation [4] and so on. Here for simplicity, we assume the use of the equivalence transformation from C to D. In the process of information granulation, we put all the numbers associated with a word *w* in a single cluster. Therefore, each word becomes a word granule having a basic representation. For instance, the basic representation for the *ith* word granule is of the form: $w_i = \{v_{i1}, v_{i2}, \ldots, v_{ij}, \ldots\}$, where v_{ij} is the weight of w_i in the *jth* page.

Now, through the information abstraction process, we can convert both word granules and page granules from their basic representation to a corresponding composite representation. Hence, a page granule is represented as a composition of word granules; by dual view, a world granule is represented as a composition of page granules. For this example, we use fuzzification for information abstraction. The fuzzy membership value for a word granule belonging to a page granule may be calculated as $\mu(w_i, p_j) = \dfrac{v_{ij}}{\max_k(v_{ik})}$. Again, by dual view, we have $\mu(p_j, w_i,) = \mu(w_i, p_j)$.

By computing with granules, we can generate various relationships among both word granules and page granules based on fuzzy set operations. For instance, the following relations can be derived among word granules:

Correlation: $w_i \leftrightarrow w_j$, if and only if

$$\frac{w_i \cap w_j}{w_i} > th_1 \wedge \frac{w_i \cap w_j}{w_j} > th_1$$

Specification: $w_i \rightarrow w_j$, if and only if

$$\frac{w_i \cap w_j}{w_i} > th_1 \wedge \frac{w_i \cap w_j}{w_j} < th_2$$

Subcategory: $w_{i1} + w_{i2} + \ldots w_{ik} \ldots = w_j$, if and only if $w_{ik} \rightarrow w_j$ for every *k* and $w_{i1} \cup w_{i2} \cup \ldots w_{ik} \ldots \leftrightarrow w_j$

V. GRANULAR MODELING OF COMPUTING WITH WORDS

The granular computing model for information analysis (figure 1) suggests that perception based knowledge can be expressed as granulated, abstracted, and semantic enriched numeric based information. Therefore, granular computing may be used to implement perception based computing [2,3,9] through numeric computing processes performed by computers. As illustrated by the dotted lines in figure 1, given a query specified in a natural language, the knowledge base will be accessed first to determine if any exiting knowledge can directly address the query. Exiting knowledge (including both common sense and domain knowledge) is the result of past granular computing processes. If no direct answer or only a partial answer is obtained from the knowledge base, then the granular modeling process is conducted to model the query as an overall granule. Next, supporting data may be needed to decompose the overall granule into groups of granules at proper levels of granularity. Subsequently, if there are no suitable computing-with-granules methods that can be directly applied to the decomposed granules, information extraction is needed to convert the composite representation to a corresponding set of basic representations, following by numerical transformations, granulations, information abstraction, and then computing with granules. The semantic enriched results are finally presented to the user and added to the knowledge base.

VI. CONCLUSIONS

In this paper, we use a set of unified components to conduct granular modeling for problem solving paradigms in several fields of computing. Each identified component may represent a potential research direction in the field of granular computing. A granular computing model for information analysis is proposed. The model may suggest that granular computing is an instrument for implementing perception based computing based on numeric computing. In addition, a novel granular language modeling technique is proposed for information extraction from web pages. This paper also suggests that the study of data mining in the framework of granular computing may address the issues of interpretability and usage of discovered patterns.

REFERENCES

[1] J. Ponte & W. B. Croft, "Alanguage-modeling approach to information retrieval." *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 275–281, 1998

[2] L.A. Zadeh, "Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/ intelligent systems", *Soft Computing*, 2, pp. 23-25, 1998.

[3] L. A. Zadeh, "From computing with numbers to computing with words — from manipulation of measurements to manipulation of perceptions". *IEEE Trans. Circ. Syst.*, Vol. 45, No. 1, pp. 105–119, 1999.

[4] S. K. M., Wong, W. Ziarko, , & P.C.N., Wong, "Generalized vector spaces model in information retrieval." *Proceedings of the 8th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 18–25. 1985.

[5] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas & R. A. Harshman, "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*, 41, pp. 391–407, 1990

[6] T. Y. Lin, "Granular computing of binary relations I: data mining and neighborhood systems." Rough Sets and Knowledge Discovery, edited by Polkowski and Skowron, Physica-Verlag, pp 107-121, 1998.

[7] T. Y. Lin, "Granular computing of binary relations II: rough set representation and belief functions." Rough Sets and Knowledge Discovery, edited by Polkowski and Skowron, Physica-Verlag, pp. 121-140, 1998.

[8] T. Y. Lin, "A roadmap from rough set theory to granular computing", *Proceeding of Rough Set and Knowledge Discovery*, pp. 33-41, 2006.

[9] T. Y. Lin, "Granular computing: from rough sets and neighbourhood systems to information granulation and computing in words." *European Congress on Intelligent Techniques and Soft Computing*, pp. 1602-1606, 1997.

[10] Y.Y. Yao, "The art of granular computing," *Proceeding of the International Conference on Rough Sets and Emerging Intelligent Systems Paradigms,* LNAI 4585, pp. 101-112, 2007.

[11] Y.Y. Yao, C. J. Liau, and N. Zhong, "Granular computing based on rough sets, quotient space theory, and belief functions," *Foundations of Intelligent Systems*, LNAI 2871, Zhong, N. Ras, Z.W., Tsumoto, S. and Suzuki, E. (Eds), The 14th International Symposium, pp. 152-159, 2003.

[12] Y.Y. Yao, "Information granulation and rough set approximation," *International Journal of Intelligent Systems*, Vol. 16, No. 1, 87-104, 2001.

[13] Y.Y. Yao, Granular computing for data mining, *Proceedings of SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security,* Dasarathy, B.V. (Ed.), pp. 1-12, 2006.

[14] Y.Y. Yao, "Modeling data mining with granular computing," *Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC 2001)*, pp. 638-643, 2001.

[15] Y.Y. Yao and B. Zhou, "A logic language of granular computing," *Proceedings 6th IEEE International Conference on Cognitive Informatics*, pp. 178-185, 2006.

[16] Y.Y. Yao Rough sets, neighborhood systems, and granular computing, *Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 1553-1558, 1999.