INTEGRATION PRINCIPLES OF RUSSIAN AND JAPANESE DATABASES ON INORGANIC MATERIALS

Nadezhda Kiselyova, Shuichi Iwata, Victor Dudarev, Ilya Prokoshev, Valentin Khorbenko, Victor Zemskov

Abstract: The methods and software for integration of databases (DBs) on inorganic material and substance properties have been developed. The information systems integration is based on known approaches combination: EII (Enterprise Information Integration) and EAI (Enterprise Application Integration). The metabase - special database that stores data on integrated DBs contents is an integrated system kernel. Proposed methods have been applied for DBs integrated system creation in the field of inorganic chemistry and materials science. Important developed integrated system feature is ability to include DBs that have been created by means of different DBMS using essentially various computer platforms: Sun (DB "Diagram") and Intel (other DBs) and diverse operating systems: Sun Solaris (DB "Diagram") and Microsoft Windows Server (other DBs).

Keywords: Databases integration, metabase, distributed information system, inorganic substances and materials, EII, EAI.

ACM Classification Keywords: H.2.4 Distributed databases, H.2.8 Scientific databases, J.2 Chemistry.

Introduction

At present rich variety of databases on properties of inorganic substances and materials were developed and maintained in the world [Bale and Eriksson, 1990; Dudarev et al., 2006; Eriguchi and Shimura, 1990; Khristoforov et al., 2001; Kiselyova, 2005; Kiselyova et al., 2004, 2005, 2006; Villars et al., 2004; Xu et al., 2006; Zemskov et al., 1998]. Traditional areas, that DBs cover, are thermodynamic, thermo-chemical, crystallographic and crystal chemical properties. The majority of large industrial corporations support DBs developments that contain the information on physical, technical and technological parameters of materials and substances. The development tendencies of modern DBs on inorganic substances and materials properties are following:

1. Internet-access to the information.

2. Powerful DBMS usage: Microsoft SQL Server, Oracle, IBM DB2, etc.

3. Great attention has been concentrated on stored information quality (reliability). Highly skilled specialists are engaged in development process of the most "advanced" commercial information systems for data capture and data reliability expert estimation. So users receive not simply "row" information but recommended values passed filtration for misprints elimination.

4. Often DBs are supplemented with information analysis tools: from traditional thermodynamic calculations and statistical procedures up to modern means for regularities search in the data allowing predicting objects behavior and making decisions. In the last case usual DBs, oriented to transaction processing, are often supplemented, for example, with special integrated information systems, that are known in English literature as *Data Warehouse* [Kimball and Caserta, 2004]. They are intended for data coordination and integration from various information sources and its preparation for subsequent computer analysis.

5. DBs on substances and materials properties integration. In this case user can find the most complete cumulative information on certain substance properties.

The last problem, resources on inorganic substance and material properties information integration is the most important today. The data on various properties of a certain substance or material are distributed among different heterogeneous DBs. The chemist or material scientist has to look through a great number of DBs in order to find necessary information. Therefore some superstructure above DBs, that will allow to output some cumulative – the integrated information on all properties set of a substance stored in different information systems, is required. That is, DBs integration is necessary. This problem solution is concerned with several difficulties. Databases on inorganic substance and material properties have been developed in various organizations and countries and thus they use different database management and operating systems. Taking into consideration differences in data quality, data expertise procedures, data formats, languages and many other troubles it should be stated that

full and smooth information resources integration is practically impossible problem. We have developed an approach to DBs integration taking into consideration DBs on inorganic substance and material properties peculiarities. The approach can be used for Russian and Japanese DBs integration in this knowledge domain.

Known Approaches to Database Integration

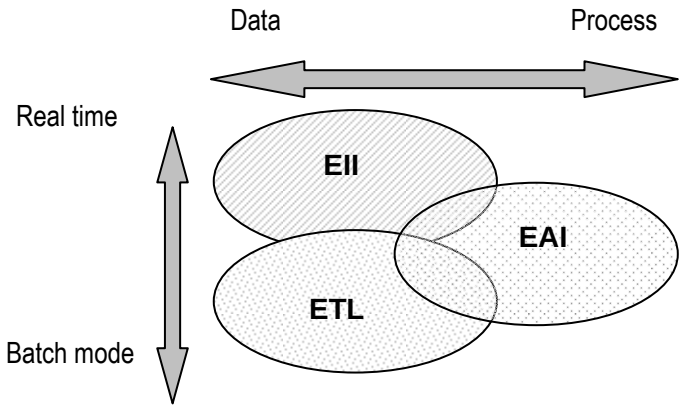

Fig. 1. Modern approaches for information systems integration.

Principally there are three approaches to database integration [Imhoff, 2005]:

1) Data Warehouse based on ETL (Extract, Transform, Load) paradigm [Kimball and Caserta, 2004].

2) EII (Enterprise Information Integration) [Morgenthal, 2005].

3) EAI (Enterprise Application Integration) [Morgenthal, 2000].

These approaches can be used to solve wide set of problems: from real-time integration to batch integration and from data integration to applications integration. Fig. 1 illustrates these approaches application area in relation to different task types [Imhoff, 2005]. The EII technology is the best approach for real-time data integration. The ETL technology allows the best batch data integration. The EAI technology gives the best results at applications integration in real-time or batch modes.

The ETL-technology implies existing resources full merging (fig. 2). That is the case when database complex is a single information system (*megabase*) for end users, operators and administrators. This approach is also known as Data Warehouse [Imhoff, 2005; Kimball and Caserta, 2004]. So at first information is extracted from DBs to be integrated. Then these data are somehow processed for clearing (that is, check for discrepancies and obviously false data elimination) and transformations – series of special procedures that allow to get a common unified format and scale. Only after these stages cleared and unified data are input into data warehouse or megabase. Database exploitation costs reduction and information duplication reduction can be mentioned among this integration approach advantages.

The second integration approach is based on EII-technology (fig. 3). It is not going to integrate databases themselves [Imhoff, 2005; Morgenthal, 2005]. Integrated data are not transferred into a central megabase but remain in the same information systems, as before. Instead the program interface for data access is developed that allows retrieving required data. EII is data integration means from multiple systems into a unified, consistent and accurate representation format geared toward the data manipulation and browsing. So the data are aggregated, restructured and relabeled (if it is necessary) and presented to a user. Usually the result of this approach is a virtually integrated heterogeneous distributed information system.

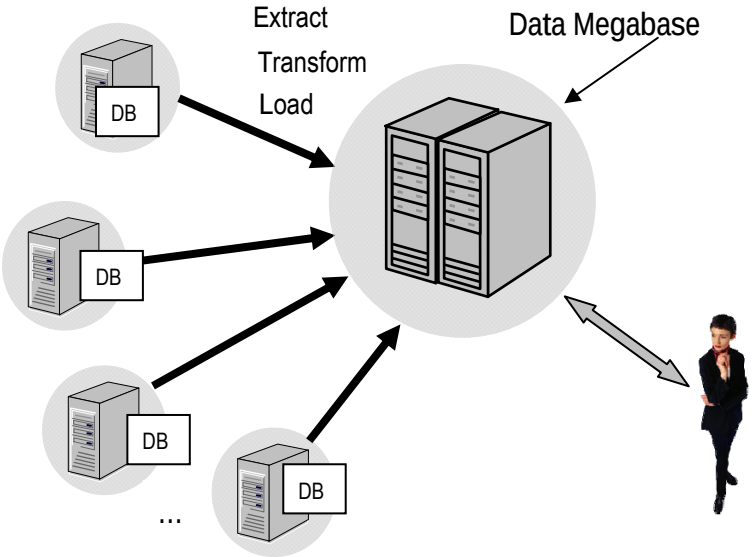

Fig. 2. ETL-approach – existing DBs full merging.

The third approach – EAI – (fig. 4) is aimed for applications integration [Imhoff, 2005; Morgenthal, 2000]. Integration can be carried out in batch or real-time mode. Combined work of two and more applications can be achieved using this approach. This approach is based on message exchange between several applications. Frequently such information exchange is carried out through some common message exchange infrastructure known as message bus. Applications are connected to this common message bus by means of special program adapters.

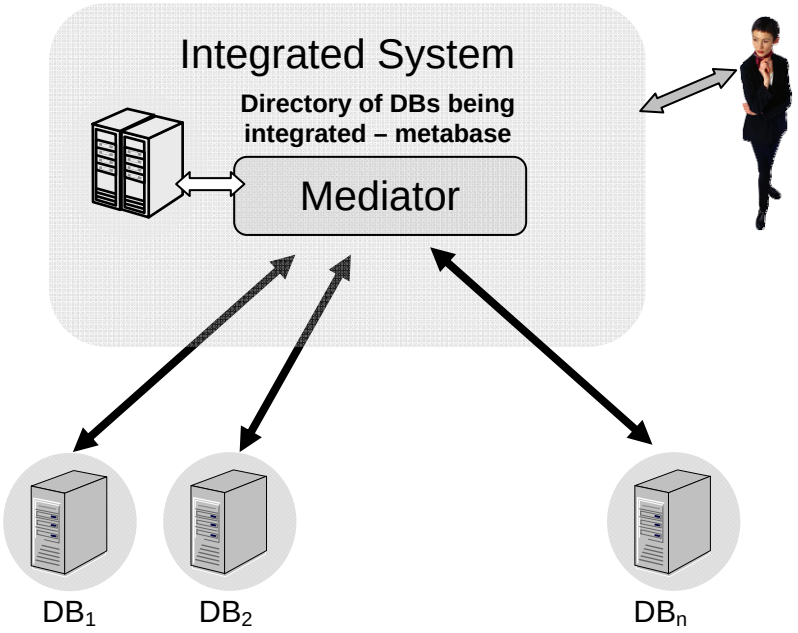

Fig. 3. EII-approach – real time data integration.

The EII and EAI technologies allow not to change every integrated database structure dramatically (and thus established database administration technology). So called "virtual" database integration and heterogeneous distributed information system creation implies independence in evolution of separate subsystems and at the same time allows to end user to get access to the whole "live" data array on a certain chemical substance or material that is stored in databases of virtually united system.

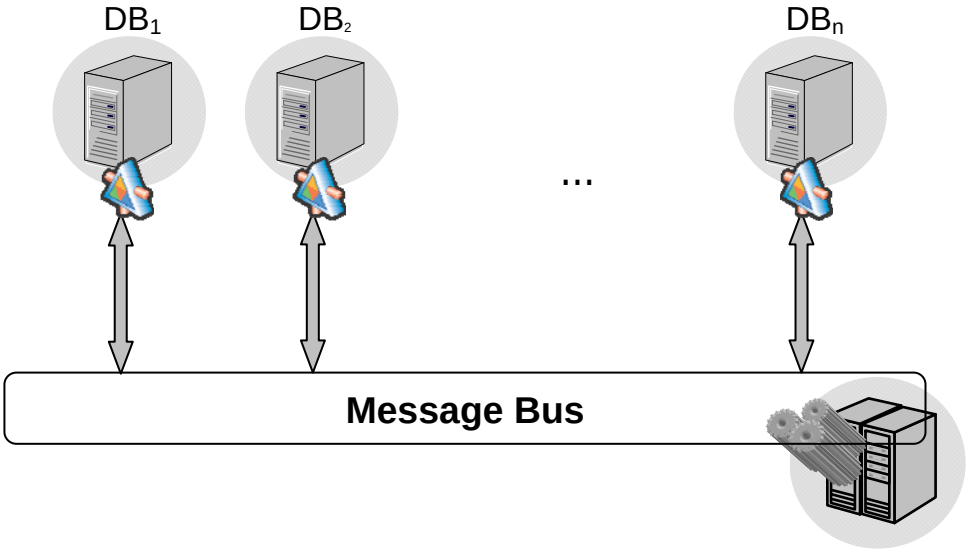

Fig.4. EAI-approach – applications integration.

So EAI technology integrates transactions of two or more applications, ETL technology merges the data of several information sources into a single one, and EII technology carries out virtual data integration of various information sources. It should be mentioned that no approach can solve all tasks arising when integrating information systems on material and substance properties.

It is necessary to take into consideration that every data center on materials properties is a point of information concentration and data analytical processing based on different software and hardware. The technology of

information accumulation and data processing has been settled down in each organization. So, great investments that have been made in hardware and software do not allow mechanically transporting all the data into some centralized database. Moreover many DBs on material and substance properties are equipped with ancillary programs for substance parameters calculation. Therefore taking into consideration current development conditions of databases on inorganic substance and material properties the integrated system based on both EAI- and EII-technologies has been developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006] (fig. 5). It allows dynamically integrate a plenty of heterogeneous databases that are supplied with any computational subsystems.

Integration of Russian Databases on Inorganic Material and Substance Properties

From the beginning the proposed approach has been used for integration of Russian DBs on inorganic material and substance properties. For successful integration solution it is needed some coordinating center, which "knows" what information is stored in every integrated DB. Such function can be carried out by *metabase* – a special metadata database that stores information on integrated DBs contents, namely, about chemical systems, substances and its modifications. Every chemical system is identified by a set of chemical elements, which are included into its composition. Each chemical substance is determined by a set of chemical elements (as a system) and their quantitative composition in the substance. Every chemical modification is defined as chemical substance having special crystal structure of phases. Metabase contains also information on properties, which data are stored in different DBs, and other data. This information is enough to make search for relevant chemical systems and data on substances and materials properties.

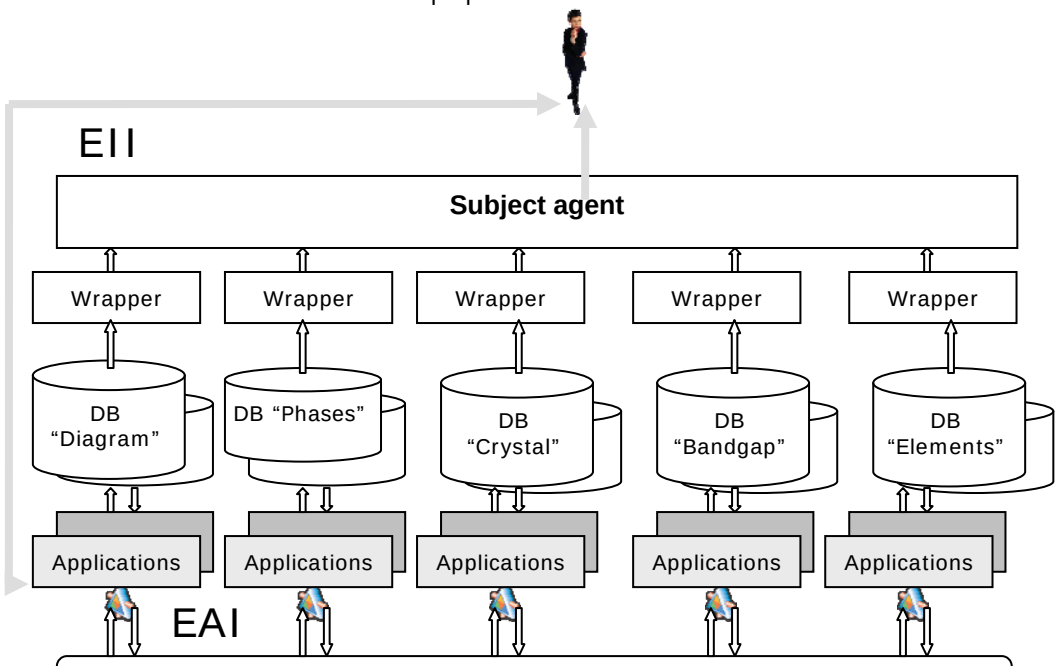

Fig. 5. Structure of integrated DBs system of Baikov Institute.

Currently the integrated information system includes five DBs that have been developed by Baikov Institute: DB on inorganic compounds properties "Phases" [Kiselyova et al., 2006], DB on semiconducting systems phase diagrams "Diagram" [Khristoforov et al., 2001], DB on substances with significant acousto-optical, electro-optical and nonlinear-optical properties "Crystal" [Kiselyova et al., 2004], DB on inorganic substances forbidden zone width "Bandgap" [Dudarev et al., 2006] and DB on chemical elements properties "Elements" (fig. 5). One of the most important developed integrated system features is that DBs which have been included into integrated system have been created with various DBMS using essentially different computer platforms: Sun (DB "Diagram") and Intel (other DBs) and different operational systems: Sun Solaris (DB "Diagram") and Microsoft Windows 2003 Server (other DBs). However the way, offered by us, has appeared successful even in such a difficult case for program realization.

Integration of Russian and Japanese Databases on Inorganic Material and Substance Properties

Next stage is an integrated system expansion. Baikov Institute information system will be integrated with other Russian [Zemskov et al., 1998] and foreign DBs [Villars et al., 2004; Xu et al., 2006] on inorganic materials and substances. Integration principles are based on the application of metabase and combined approach that has been developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006]. Sometimes small additional tables, that contain information about elements sets and their contents in substance and crystal structure, should be included into these DBs.

The following metabase structure can be used for Web-applications integration of DBs on inorganic substances and materials properties (fig. 6).

Tables designation (fig. 6): DBInfo – main table containing information on DBs Web-applications to be integrated; UsersInfo, UsersAccess - tables containing information on integrated system users and their access permissions to information; SystemInfo, PropertiesInfo, DBContent – tables that describe integrated resources contents (what information on chemical systems and their properties is stored in what DB); CompatibilityClasses, Compatibility, Systems2ConsiderInCompatibility – tables that contain information on relevance classes and determine relevant chemical systems.

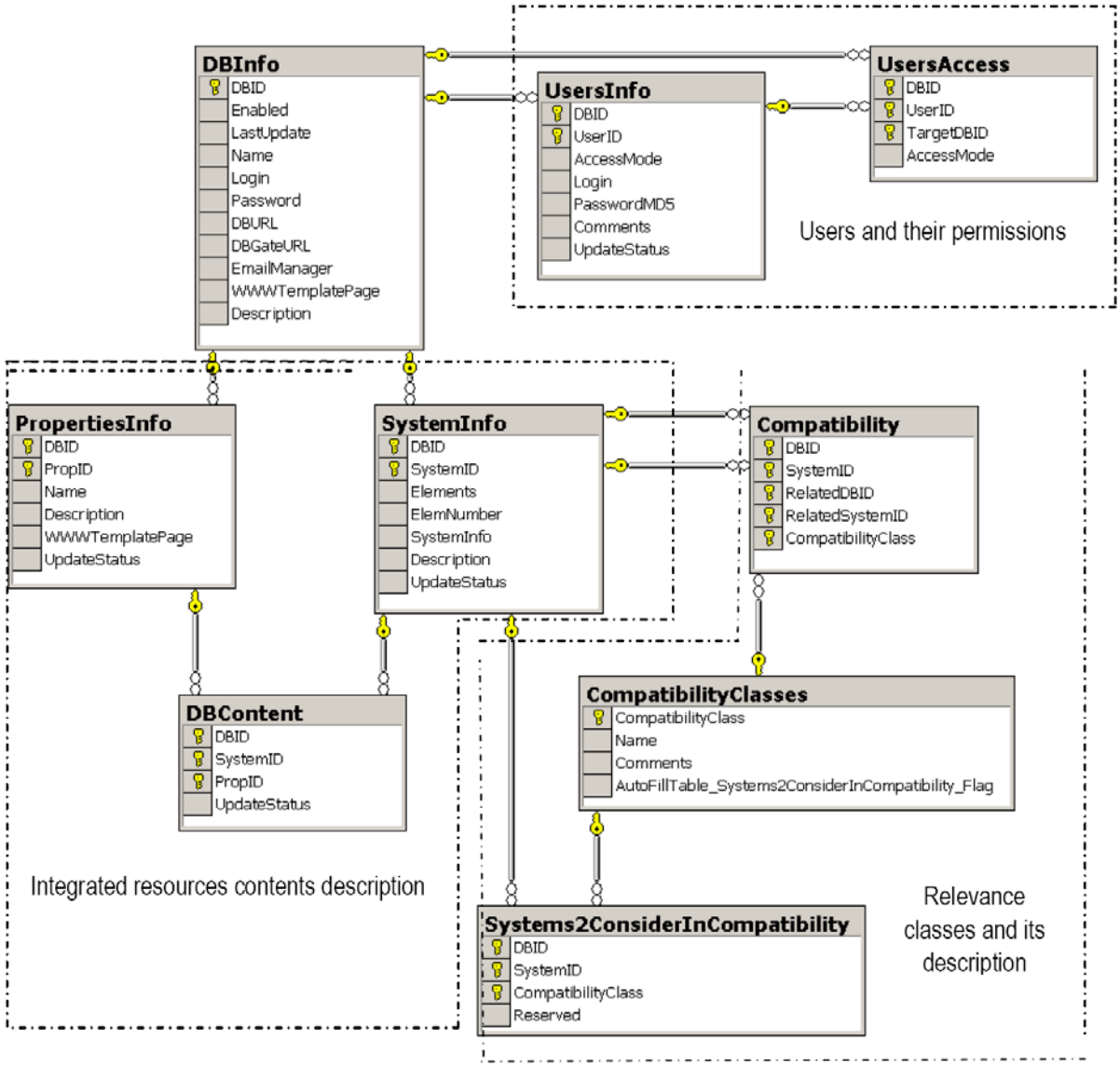

Fig. 6. Metabase structure for DBs Web-applications integration.

Conclusion

The complex approach to information integration combining integration at data level and at user interfaces level (EII+EAI) is offered. Within proposed approach access means have been implemented to all current user interfaces of virtually united information system. Moreover the system allows users to move transparently between different applications (EAI). According to the common developed information schema subject mediator has been implemented. It provides rich opportunities for information extraction and aggregation from diverse distributed data sources on material and substance properties (EII).

Search for relevant data in integrated information system tasks and transparent user transition between DBs Web-applications implementation (taking into account the security issues) have been solved during DBs Web-applications integration. Metadata database (metabase) has been used for relevant information search mechanisms implementation. Matabase is a special reference database containing metadata only. Metadata are information on information systems to be integrated. Diverse data sources integration is based on conceptual knowledge domain structure (inorganic chemistry) and heterogeneity conflicts resolution ways development.

Databases on inorganic material and substance properties system is accessible for registered users via Internet: http://www.imet-db.ru.

The work is supported by RFBR, grants №06-07-89120 and 05-03-39009.

Bibliography

[Bale and Eriksson, 1990] C.W.Bale and G.Eriksson. Metallurgical thermochemical databases a review. Can.Met.Quart. 1990, v.29.

[Dudarev et al., 2006] V.A.Dudarev, N.N.Kiselyova, V.S.Zemskov. Integrated system of databases on properties of materials for electronics. Perspektivnye Materialy, 2006, N.5 (Russ.).

[Eriguchi and Shimura, 1990] K.Eriguchi and K.Shimura. Factual databases for materials design and manufacturing. ISIJ Int., 1990, v.30.

[Imhoff, 2005] C.Imhoff. Intelligent Solutions: Understanding the Three E's of Integration EAI, EII and ETL. DM Review Magazine, 2005, apr. (http://www.dmreview.com/article_sub.cfm?articleId=1023893).

[Khristoforov et al., 2001] Yu.I.Khristoforov, V.V.Khorbenko, N.N.Kiselyova, et al. Internet-accessible database on phase diagrams of semiconductor systems. Izvestiya VUZov. Materialy elektron.tekhniki, 2001, №4 (Russ.).

[Kimball and Caserta, 2004] R.Kimball and J.Caserta. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004.

[Kiselyova, 2005] N.N.Kiselyova. Computer Design of Inorganic Compounds. Application of Databases and Artificial Intelligence. Nauka, Moscow, 2005 (Russ.).

[Kiselyova et al., 2005] N.N.Kiselyova, V.A.Dudarev, I.V.Prokoshev, et al. The distributed system of databases on properties of inorganic substances and materials. Int.J."Information Theories & Applications", 2005, v.12.

[Kiselyova et al., 2006] N.Kiselyova, D.Murat, A.Stolyarenko, et al. Database on ternary inorganic compound properties "Phases" in Internet. Informazionnye resursy Rossii, 2006, N.4 (Russ.).

[Kiselyova et al., 2004] N.N.Kiselyova, I.V.Prokoshev, V.A.Dudarev, et al. Internet-accessible electronic materials database system. Inorganic materials, 2004, v.42, №3.

[Kornuyshko and Dudarev, 2006] V.Kornuyshko and V.Dudarev. Software Development for Distributed System of Russian Databases on Electronics Materials. Int. J. "Information Theories & Applications", 2006, v.13.

[Morgenthal, 2000] J.P.Morgenthal. Enterprise Applications Integration with XML and Java. Prentice Hall PTR; Bk&CD Rom edition, 2000.

[Morgenthal, 2005] J.P.Morgenthal. Enterprise Information Integration: A Pragmatic Approach. Lulu.com, 2005.

[Villars et al., 2004] P.Villars, M.Berndt, K.Brandenburg, et al. The Pauling File, binaries edition. J.Alloys and Compounds, 2004, v.367.

[Xu et al., 2006] Y.Xu, M.Yamazaki, H.Wang, K.Yagi. Development of an Internet system for composite design and thermophysical property prediction. Mater.Trans., 2006, v.47.

[Zemskov et al., 1998] V.S.Zemskov, F.A.Kuznetsov, V.B.Ufimtsev. Databanks on semiconducting and other materials for electronics and technology of their productions. Izvestiya VUZov. Materialy elektronnoi tekhniki, 1998, N.3 (Russ.).

Authors' Information

Nadezhda Kiselyova – *A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: kis@ultra.imet.ac.ru*

Shuichi Iwata – *Graduate School of Frontier Sciences, The University of Tokyo, P.O.Box: 113-8656, Room 507A, Building E12 QUEST, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, e-mail: iwata@k.u-tokyo.ac.jp*

Victor Dudarev – *A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: vic@osg.ru*

Ilya Prokoshev – *A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: eldream@e-music.ru*

Valentin Khorbenko – *A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: Khorbenko_v@mail.ru*

Victor Zemskov – *A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: zemskov@ultra.imet.ac.ru*

IMAGE QUOTIENT SET TRANSFORMS IN SEGMENTATION PROBLEMS

Dmitry Kinoshenko, Sergey Mashtalir, Konstantin Shcherbinin, Elena Yegorova

Abstract: Image content interpretation is much dependent on segmentations efficiency. Requirements for the image recognition applications lead to a nessesity to create models of new type, which will provide some adaptation between law-level image processing, when images are segmented into disjoint regions and features are extracted from each region, and high-level analysis, using obtained set of all features for making decisions. Such analysis requires some a priori information, measurable region properties, heuristics, and plausibility of computational inference. Sometimes to produce reliable true conclusion simultaneous processing of several partitions is desired. In this paper a set of operations with obtained image segmentation and a nested partitions metric are introduced.

Keywords: image, spatial reasoning, partitions, covers, interpretation.

ACM Classification Keywords: I.4.6 Segmentation: region growing, partitioning

Introduction

Modern phase of developing intellectual systems for information processing in correlation-extremal tracking, industry robotics vision, graphical and graphological information processing, medical diagnostic complexes, etc. requires ability to process different visual data for its unsupervised context interpretation. Increasing of arbitrary image identification reliability in real time necessitates refinement of complex images recognition under uncertainty factors.

Efficiency of image structuring and understanding strongly depends on a segmentation as a process of separating an image into several disjoint (or weakly intersecting) regions whose characteristics such as intensity, color, texture, shape, etc. are similar [see e.g. 1-4]. Segmentation is a key step in early vision and it has been widely investigated in image processing. Generally this process is rather laborious and not completely algorithmized for arbitrary images. Different data registration conditions, by-product facts, lack of robustness for the disturbing effects – this is a far not complete list of the reasons, which refers the process of image recognition to the class of not ordinary tasks. In practical applications the choice of methods which are able to form the most accurate regions of interest is of the prior importance. Unfortunately most of existing methods produce only the primary partitions which can not guarantee adequate image interpretation as image content formal descriptions obtained by using only low-level features are not necessarily the case for true conclusions. We may get totally correct segmentation, but in most cases we obtain under-segmentation, over-segmentation, missed regions, and