2009 Fourth International Conference on Internet Computing for Science and Engineering

Publish/Subscribe Driven Matching Algorithm

Min Song[1,2], Zhe Zhang[1,3], Guisheng Yin[1],
[1]College of Computer Science and Technology
Harbin Engineering University
Harbin, China
soongmin@163.com
[2]Beijing Foreign Studies University
Beijing, China

Hongbin Wang[1], Jun Ni[4]
[3] System Engineering Research Institute
Beijing, China
[4] Department of Computer Science
University of Iowa
Iowa City, IA 52242, USA

***Abstract*—The characteristics of publish/subscribe technology are of asynchronism and loosely coupled. This phenomenon makes producers and consumers are de-communicated in temporal and spatial domains during processing. It requires develop a dynamic integration in today's large-scale, distributed information system. This paper introduces a Semantic Integration Data Model (SIDM) for publish/subscribe systems, within which model subscription module is designed. The model contains a matching algorithm which combines both publishing-driven and subscription-driven matching algorithms. Our experiments show that this matching algorithm has high efficiency for matching.**

Keywords- Publish/subscribe, information system, integration, matching, algorithm

I. INTRODUCTION

With the rapid development of Internet applications including mobile, grid, pervasive computing, it become realistic to quickly and accurately obtain desired information in dynamic, heterogeneous, and open environment. The large-scale distributed systems being developed significantly meet the unprecedented need. However, it is often required to have a more flexible communication model and interactive mechanism to deal with a highly-dynamic and loosely-coupled system characteristics. This demand lies in today publish/subscribe (pub/sub) systems, in which information providers provide publishing services, while subscribers (authors for example) provide information contents. Both requests can be considered as events. For a special domain theme, pub/sub providers confirm their services to solicit useful topics for publications, to ensure technical publication process, and to schedule publication workflow on time. On the other hand, publisher's events strongly rely on subscribers' intents such as interests, writing timeframe, technical writing, information passing, intellectuality, copyright, etc. The virtue of pub/sub communication is based on the information exchange between the producer and consumer, especially in terms of time, space (geologically), and necessary processes work flow.

The technology to ensure an event within a pub/sub system can be a middleware. The middleware features the interactivity and interoperability among distributed computing platforms. In a publish/subscribe system, the interactive information between producers and consumers can be considered as an event. The can be classified into two groups: theme-based and content-based systems. In theme-based system (such as IBM's MQSeries[4]), each event was divided into a number of fixed themes. Each event belongs to a specific theme. A publisher must specify their domain themes in advance, while subscribers can subscribe their contents to the particular publishers. The subscriptions can also be considered as events. In the content-based system, the publisher, in accordance with the events internally designed, sets up a subscription mechanism with conditions. Then, all the events to meet the conditions will be transmitted to the subscribers. Compared with the theme-based pub/sub systems, the content-based pub/sub system provides a large.

The current content-based pub/sub system has two models. One is map-based and the other is XML-based. In the map-based pub/sub system, the content of event has a set of "attribute=value". The subscription condition is generally simple assertion of connection based on various attributes, referred to as the flat pattern. Existing influential prototype systems include SIENA[5], Gryphon[6], JEDI[7], Keryx[8], and Elvin[9]. In XML-based pub/sub systems, each event is described by an XML document. The subscription language usually is Xpath or its variants (including the restraints of XML document structure and certain elements and attributes. It usually expressed in usually hierarchic structure. Such prototypes can be found in XFilter[10], XTrie[11], and WebFilter[12].

The core of publish/subscribe system is a subscription matching algorithm. It is related to publisher's organizational structure that directly affects matching efficiency. The matching algorithm is also related to the matching process of subscribing. It directly affects the matching success. In publish/subscribe system, due to the lack of a unified description, manifestation, and operation model of publication, the publisher's organizational efficiency becomes low. That is because most publish/subscribe systems support complex semantics. The current matching process uses publish-driven algorithm that makes the event subscribers lost their access ability to update information.

In order to solve the above problem, we apply a data model with semantic space integration to the pub/sub systems. The ultimate goal is integrate the information at different resources interactive and dynamically. Combined with the characteristics of information system integration and

978-0-7695-4027-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICSE.2009.49

IEEE computer society

the features of pub/sub system, we propose double-side driven matching algorithm for pub/sub systems.

II. SPACE INTEGRATION DATA MODEL (SIDM)

OIM-based of Southeast University, Harbin Engineering University proposes Space Integration Data Model [13]. In the data model, data objects can be divided into point structure, line structure, region structure and constructing structure. The space object of data structure model is showed as basic data type object and complex data type object, that are defined as follows:

α: 0-dimension space object set, a:0-dimension space object element, that is, atom data type, including NUMBER, REAL, INEGER, STRING, BOOLEAN, BINARY, etc.;

β: 1-dimension space object set, ß:1-dimension space object element, that is, simple data type. It is modified based on atom data type. All the key words on the upper and lower edges have different rules and restrictions, these restrictions are implicit in the semantic constraints of space data structure. Using explicit expression constraint semantic to ensure the semantic consistency during the exchange of product data;

γ: 2-dimension space object set, ν:2-dimension space object element, that is, integrate data type. It is modified based on simple data type, with the corresponding key words, is a line set of simple data type;

3: 3-dimension space object set, ɜ:3-dimension space object element, that is, constructing data type, it is the data type that constructed with a specific semantic belonging to different meta-statement atom data type, simple data type and integrate data type.

Space data model has the following relationship:

ß ⊇α={a1,a2, …};

ν ⊇β={α1,α2, …};

ɜ ⊇ (γ∨β∨α) = {{ν1,ν2,…} ∨ {ß1,ß2,… }∨ {a1,a2, …}}.

The data object type of data model with semantic has atom data type, simple data type, integrate data type and constructing data type. The object attributed to different types, then the describer of object is different too, the concrete grammar described as:

SIDM: Atom | Simpleness | Integration | Construction;

Atom: <subject, oid, type, value>.

Simpleness: <subject, oid, type, a-oid-list>.

Integration: <subject, oid, type, s-oid-list>.

Construction: <subject, oid, type, ref-oid-list>.

Oid is the single object identifier; subject shows the meaning that object represented, type is object type, to the atom object, it is the atom data type that model allows, to the complex object, type is the set data type; value is the atom type value, the complex object value of a-oid-list, s-oid-list and ref-oid-list for a list of <rank, oid>,for the sub-object set what the object include, that is, object citing, in which a-oid-list has rank::0 | 1 , rank have the value 0 indicates that sub-object is atom type, rank have the value 1 indicates the series of atom type; s-oid-list has rank::0 | 1 | 2 , rank have the value 0 indicates sub-object is atom type, rank have the values 1 indicates the series of atom type, rank have the values 2 indicates the line series of simple type; ref-oid-list has rank::0 | 1 | 2 | 3 | 4,rank have the value 0 indicates sub-object is atom type, rank have the value 1 indicates the series of atom type, rank have the value 2 indicates the line series of simple type, rank have the value 3 indicates the series of integrate data type, rank have the value 4 indicates the combination of series of simple type (atom type) and integrate data type. According to the Construction described, the type that ref-oid-list quote object for inner citing quotation and outer quotation. Outer quotation includes simple object quotation and extended object quotation. Therefore, SIDM defines three quotation types: inner, simple and extended, they are show inner quotation, simple quotation and extended quotation. Quoting type for the quotation object of inner, its value is oid of quotation object, quoting type for object value of simple and extend is like value of complex object, for a group oid list manifest this quotation object contains sub-element, attribute and characteristics data.

III. INFORMATION SYSTEM INTEGRATION-ORIENTED PUB/SUB MODEL, ISIOPM

The main characteristic of distributed information system is the information structure of system is relatively stable. And the data of information is variable in the information structure framework. SIDM and ask pub/sub model well adapted to the feature of distributed information system. Most of the current publish/subscribe systems [10] [11] [12] will be the information structure and information value of data publishing and subscribing by the form of XML. This will has greatly influenced for the efficiency of information publishing and efficiency of information subscribing to match. To overcome these problems, pub/sub model of information system integration-oriented separate the publish of information structure from publish of information data value. Added the time characteristics in subscribe operation of information, this structure of information only publish once. And information value of data can be sent to pub/sub system continuously. At the same time, information subscribes only once, which information satisfying the conditions of subscription will be continuously returned to the subscribers.

To provide a general way of information for distribution, heterogeneous and dynamic information. ISIOPM take a unified data model, which describes the data model to by information describer. The describer will be responsible for the application semantic information to show the single information expression structure of system explaining of . System to take advantage of semantic information of event to match the structure of information, which can provide more precise, filtering than key words, based method. Figure 1 describes the information exchange model of pub/sub model of information system integration-oriented.

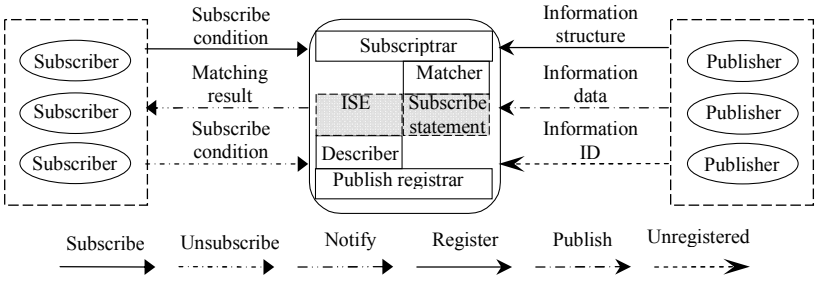

Figure 1. ISIOPM concept model

Information producers (publishers) register the information structure with Publish registrar by event form and obtain the information identifier, and send information identifier with specific data information to Publish registrar continuously. If the information producer is no longer make information, or the information has no effect, then write-off the information to Publish registrar. Consumers (subscriber) send a subscription to Subscriptrar, said the system in which information is interesting the subscribers with the time characteristics. Information that satisfy the subscription back to subscribers in real-time or cycle mode; if no longer interested in, it can cancel subscribe. According to the SIDM model of ISIOPM, describer uniform describe and convent the information that publisher published and it set information structure entity for it's publish information. According to the subscribe statement, matcher is responsible for efficiently find the subscription condition to match the matching result is given and send information to subscribers.

IV. SUPPORT INFORMATION INTEGRATION SUBSCRIBE STATEMENT

In order to achieve a highly efficient information subscription matching, it must have the support of subscribe statement. Event must comply with SIDM, so the user's subscribe condition is actually a subscribe statement model based on SIMD grammar which provides the restriction of subscription.

In ISIOPM, a number of "statement pattern" of "and" compose user's a subscription condition. Each statement pattern describe a statement of subscription, it's form as follows:

(subject, meta-statement,[filter_func(subject)])

"Subject" prescribes a statement of "subject", they are mark the concrete information object, meta-statement prescribes a statement should meet the type of constraint, which is make a certain statement pattern meta-statement for (s, p, r). If a certain SIDM statement can match with the statement pattern, the following assertions are true:

srdf: typeoid-list-subject

(s must belong to a certain specific value of oid-list a)

prdfs: subPropertyOfoid-list-Atom-subject

(p must be a certain Atom of subject type)

rrdf: typeoid-list-Atom-value

(the condition that limits must belong to the value of Atom)

In statement pattern, subject is variable and it type for Atom, in the statement pattern can has a filter function filter_func(subject), that is Boolean expression to limit the values of object variables further. Filter function allows relational operation, such as >, <, =.

V. DOUBLE SIDE DRIVEN MATCHING ALGORITHM

A. Overview

At present, a large number of subscribe conditions are saved and set index in general pub/sub systems, so that when come to a publish can find matching subscribe condition rapidly. Now this matching way and algorithm are not suitable for integration with information systems. In the information system integration, whatever the subscription or data publish achieved is need return the system that all meet the information of subscribe condition to subscriber. But now the matching method and algorithm existed are can not return the system existed information to new reached to meet the information subscriber of subscribe condition. A double side driven matching algorithm is presented in this paper for adapt to the requirements of information system integration, that is publish driven matching and subscribe driven matching algorithm. The following algorithm is used to introduce the index structure and corresponding matching algorithm.

B. Algorithm Description

Maintaining all the information structure has been published and has been sent actual information data in ISIOPM system. Based on the publish information that defined by user, according to SIDM class hierarchical structure and attribute hierarchical structure, ISIOPM system based on the appropriate grammar to complete the analysis of application information structure, according to Subject organize all kinds of information structure for tree structure, said information structure tree (the abbreviation is ST), get a single information tag, and build data storage table. When information data published to system, according to the appropriate location of storage table that information tag found in the tree to being the cache of information data. Take the array table to manage all the information data, known as information storage structure array, referred to as SS. ST and

SS are composed as information structure entity, referred to as ISE. Figure 2 shows its structure. ISE is ISIOPM system the base of index structure, in which each Subject of ST by dictionary order, each Subject has an ID, in order to facilitate the search.

ISIOPM provides standard subscribe statement for users. ISIOPM get the entire legal meta-statement base on the subscribe request submitted by users. Organized them into two trees, waiting matching and matched tree. The tree root is the tag of tree, tree leaf is legal meta-statement of subscribe request. Waiting matching tree save all subscribe request and legal meta-statement which correspondent the grammar is waiting matching, and matched tree save all subscribe request and legal meta-statement have been matched.

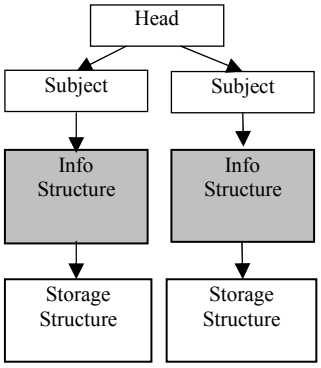

Figure 2. Information structure entity of ISIOPM system

ISIOPM take the publish driven matching and subscribe driven to match the matching algorithm of double side driven, its steps are as follows:

(1) to receive an event, if it is information publish, then execute(2), if it is subscription , then execute(3);

(2) to receive information publish, then the following steps are:

a. cache information in the ISE;
b. to find corresponding information subscribe request in waiting matching trees, if found, then execute next step, else turn o;
c. According to Subscription to find subject id in ISE, if find, then execute next step, else turn o;
d. in found the subject ID, according to the Info Structure (that is ST) of meta-statement in the ISE find the corresponding data tag items, if find, then the next step, else turn h;
e. in which ST found data tag items, according to filter_func (subject) to determine whether or not the corresponding data values are met, if the meet is the next step, otherwise turn i;
f. build meta-statement and data tag item mapping, then the next step;
g. get data , waiting for data organization, turn j;
h. mark the data tag item is failed to match in meta-statement;
i. mark the condition of data tag item is identify failed in the meta-statement;
j. if successful matching , then organize the result of g, h, i into specific information structure, and return to the subscribers, then the next step;
k. remove the Subscribe request from waiting tree, added to the tree has to matched ;
l. to search the subscribe request of corresponding information in matched tree, if find, then the next step, else o;
m. according to the establishment of meta-statement and data tag item mapping, in the corresponding data tag item, to determine whether or not the corresponding data values are met by filter_func (subject), if meet, then execute g, else turn i;
n. to match the success, then return the result of g, h, i organized into specific information structures to subscribers (Note: h is the results for the last match);
o. matching failed;

(3) Received information subscription, then the following steps:

a. according to the subject of subscription find the subject id in ISE, if found then the next step, else j;
b. in found subject ID, according to the Info structure (that is ST) of meta-statement in the ISE to find the corresponding data value tag items, if find, then the next step, else turn f;
c. in which ST found data tag items, according to filter_func (subject) to determine whether or not the corresponding data value are met, if the meet is the next step, else turn g;
d. build meta-statement and data tag item mapping, then the next step;
e. get data, waiting for data organization, turn h;
f. mark the data tag item is failed to match in meta-statement;
g. mark the condition of data tag item is identify failed in meta-statement;
h. if successful matching, then organize the result of e, f, g into specific information structure, and return to the subscribers, then the next step;
i. add the subscribe request to has been matched the tree;
j. add the subscribe request to waiting matching tree.

Double side driven matching algorithm to ensure when come a information subscribe or information publish, should matching the information what the event contained.

C. Simulation Experiment and Analysis

For the main performance evaluation of publish / subscribe matching algorithm is the time of matching. In this paper, it has simulated experiments to double side driven matching algorithm. Experimental configuration as table I.

TABLE I. SIMULATION EXPERIMENT ENVIRONMENT CONFIGURATION

Number	Item	Configuration
1	CPU	Intel Pentium IV 2.0GHz
2	memory	512M
3	OS	Windows XP
4	network	100M fast ethernet switch, net connect category 5e cable
5	programming language	C\C++

In the experiment, ISIOPM system deployed in a node, the information publisher of information are deployed in many different nodes, the same node can not only deploy subscribers but also deploy publisher.

In the simulation experiments, the assumption that in ISIOPM system information type is Construction, which the ref-oid-lis contains 10 Simpleness classes, each Simpleness class has 10 the attribute of Atom type. The assumption that each subscription has one statement pattern, each statement pattern has 5 meta-statement which point to a five Aton type "and" operation of filter function (ie filter_func). In this paper, has two aspects of experiments, one under the different subscribe number to verify the matching time of a single information type, and raised the subscribe number of information from 1000 to 5000 step by step, part of the experimental results as table II; the other is to verify the single subscribe matching time under different information number, and raised the subscribe number of information from 1000 to 10000 gradually, part of the experimental results as table III. From the experimental results, the time of subscribe matching is good.

TABLE II. A SINGLE TYPE OF INFORMATION TO MATCH THE NUMBER OF DIFFERENT SUBSCRIPTION TIME

Number	Number of information subscription	Matching time(ms)
1	1000	73.4
2	2000	102.7
3	3000	154.3
4	4000	262.8
5	5000	433.5

TABLE III. SINGLE SUBSCRIPTION IN DIFFERENT INFORMATION NUMBER

Number	number of information subscription	Matching time(ms)
1	2000	163.7
2	4000	232.9
3	6000	441.3
4	8000	668.8
5	10000	923.2

VI. CONCLUSION

Pub/sub technology is an effective method to realize information integration. This paper the application of pub/sub technology to the integration of heterogeneous information system, introduce SIDM as each unified data model of heterogeneous information system sources of information. It provides a unified, standard, scientific information standard of expression and operation. Designed information system integration-oriented pub/sub model. Put forwarded a double side driven matching algorithm, that is, publish driven matching and subscribe driven matching algorithm. ISIOP and double side driven matching algorithm have fully considered the characteristics of pub/sub technology and information system integration feature. Application of pub/sub technology to information system integration realization, to make the information producer and consumer are decoupled completely in time and space, to realize "seeing is getting" on demand obtain ability for heterogeneous information source. To improve the time efficiency and space efficiency of algorithm is the next important research.

REFERENCES

[1] Eugster PT, Felber PA, Guerraoui R, Kermarrec AM. The many faces of publish/subscribe. ACM Computing Surveys, 2003,35(2): 114-131.

[2] Lassila O, Swick RR. Resource description framework (RDF) model and syntax specification. 1999. http://www.w3.org/TR/1999/ REC-rdf-syntax-19990222/.

[3] Berners-Lee T. Using XML for data. 2001. http://www.w3.org/DesignIssues/XML-Semantics.html.

[4] IBM. Internet Application Development with MQSeries and Java. Palos Verdes: Vervante Corporate Publishing, 1997.

[5] Carzaniga A, Rosenblum DS, Wolf AL. Design and evaluation of a wide-area event notification service. ACM Trans. on Computer Systems, 2001,19(3):332-383

[6] Aguilera MK, Strom RE, Sturman DC, Astley M, Chandra TD. Matching events in a content-based subscription system. In: Proc. of the 18th ACM Symp. on Principles of Distributed Computing. New York: ACM Press, 1999. 53-61.

[7] Cugola G, Nitto ED, Fuggetta A. The JEDI event-based infrastructure and its application to the development of the OPSS WFMS. IEEE Trans. on Software Engineering, 2001,27(9):827-850.

[8] Wray M, Hawkes R. Distributed virtual environments and VRML: An event-based architecture. In: Proc. of the 7th Int'l World Wide Web Conf. (WWW7). Amsterdam: Elsevier Science Publishers, 1998. 43-51.

[9] Fitzpatrick G, Kaplan S, Mansfield T, David A, Segall B. Supporting public availability and accessibility with Elvin: Experiences and reflections. Computer Supported Cooperative Work, 2002,11(3):447-474.

[10] Altinel M, Franklin MJ. Efficient filtering of XML documents for selective dissemination of information. In: Proc. of the 26th Int'l Conf. on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2000. 53-64.

[11] Chan CY, Felber P, Garofalakis M, Rastogi R. Efficient filtering of XML documents with XPath expressions. The VLDB Journal, 2002,11(4):354-379.

[12] Pereira J, Fabret F, Llirbat F, Jacobsen HA, Shasha D. WebFilter: A high throughput XML-based publish and subscribe system. In: Proc. of the 27th Int'l .Conf. on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2001. 723-724.

[13] Zhang Zhe, Song Min, Liu Daxin, Wei Zhengxian, Song Zheng. Data Model Support Distributed System Integration[J]. Journal of System Simulation, 2009, 17(3):134-145

[14] Ma Jiangang, Huang Tao, Wang Jinling, Xu Gang, Ye Dan. Underlying Techniques for Large-Scale Distributed Computing Oriented Publish/Subscribe System[J].Journal of Software, 2006, 17(1):134-145.

[15] Xu Gang, Huang Tao, Liu Shaohua, Ye Dan. Survey on the Core Techniques of Distributed Application Integration [J]. Journal of Computer, 2005, 28(4):434-443.

[16] Luo Yingwei, Liu Xinpeng, Peng Haobo, Wang Xiaolin, Xu Zhuoqun. Information and Services Integrating and Scheduling Model for Event Handling.. Journal of Software, 2006, 17(12):2554-2564.

[17] Han Jianghong, Zheng Shuli, Wei Zhenchun, Jiang Jianwen, Wu Yongzhong. Research on Web Data Model Oriented to XML,Journal of Chinese Computer Systems. 2005,26(4):609-613.

[18] Jin Peiquan, Yue Lihua, Gong Yuchang.Smallest Set of 2-D Topological Spatial Relationship and Its Implementation in Spatiotemporal Data Model.Computer Engineering, 2004, 30(18):71-73.

[19] He Zhenying, Li Jianzhong, Wang Chaokun. A Data Model for XML Database. Journal of Software, 2006,17(4):759-769.

[20] Wang Ning , Xu Hong-bing , Wang Neng-bin. A data model and algebra for object integration based on a rooted connected directed graph[J] . Journal of Software , 1998 ,9(12) :894-898.

[21] Wang Ning , Xu Hong-bing , Wang Neng-bin. Capabilities-Based query decomposition and optimization in heterogeneous data integration system[J] . Chinese J . Computers, 1999, 22 (1):31-38.

[22] Guo Ming, Li Shanping, Dong Jinxiang, Fu Xiangjun. Research on product information model based on Ontology and Semantic Web,Journal of Zhejiang University. 2004, 38(1):22-28.