A Resource Search Model Based on Semantically Enabled P2p Grid

Jun Ye[1] Zhishu Li[1] Yanyan Ma[2] Qing Li[1]

[1] School of Computer Science, Sichuan University, Chengdu 610065,P. R. China.
[2]School of Humanity Science, Zhejiang Shuren University, ,Hangzhou 310015,P. R. China

Abstract

This paper presents a layered resource search model based on P2P Grid that has a good feature in expansibility, autonomy, modularity and loose coupling. In terms of this model a Bootstrap based on Distributed Hashing Table (DHT) is added into P2P Grid and a Peer Service is implemented in every node. The semantic-based resource management architecture and RPC protocols are discussed. A vertical search engine prototype is realized with data extraction functionality. This model could be used in P2P Grid systems and heterogeneous P2P systems as a reference for resource searching, sharing and collaboration.

Keywords: Grid, P2p, Distributed hashing table, Semantic web service, Resource search

1. Introduction

The amount of resources available on the Web has been growing explosively during the past few years. To retrieve resources including web pages, images, videos more   effciently from the Web, various approaches for querying the Web resources have been suggested, These techniques basically fall into one of two categories querying the Web with popular search engine and P2P search technology for resources. This paper investigated a resource search model by building a logical netnwork on the top of large internet web layer based on P2P and Grid technology.

The representation, discovery, access and collaboration of resources are important topics in distributed computational networks. Both Grid and P2P systems devote themselves to managing the resources for sharing in a distributed environment, although their intentions, services, infrastructures and evolvement courses are obviously different. Grid resources demand a totally decentralized, distributed, extensible and intelligent resource discovery mechanism. P2P systems are generally characterized by the lack of centralized control but high scalable. Using P2P approaches to realize decentralized P2P Grid systems extends the Grid resource scalar exponentially. It goes by the name of P2P Grid. A critical research issue in such distributed systems is how to "find any given data item in a large system in a scalable manner"[3]. This paper provides a search model with semantically enabled characteristic to solve the resources searching, sharing and collaboration problem between P2P Grid's nodes.

2. Addressed problem

The widely use of Web Service [1] technology have effectively solved the problems such as the interactive operation and information integration in the heterogeneous platform. After Grid researchers have absorbed the Web Service standard, they advance OGSA/OGSI [2] standard. OSGI makes Web Services as stateful resources. In 2004, OASIS issued WSRF [11], which had absorbed the new achievement of Web Service. WSRF distinguish the difference between Web services and stateful resources. WSRF vision has enhanced the Grid resource management in the heterogeneous platform. However, the problem of scalability in Grid still needs to be strengthened.

Using P2P approaches in resource discovery can help Grid deal with dynamic, large-scale distributed and decentralized resources in multiple autonomous administrative domains. In terms of resource search and discovery, the MDS [4] of computational Grid realized dendriform metadata catalogue service based on LDAP; UDDI of Web Service accounted for the uniform description, register and search of service entity; P2P systems constructed an overlap network based on the TCP/IP network in order to optimize resource search which has got great success. Decentralized structured topology, such as CHORD [12], CAN [13] and Kademlia [7], is mainly based on DHT and has a great potential to offer Grid scalability, reliability and maintainability. DHT systems commonly discover distributed resource as follows: each node stores a sub-space of the whole resource space, and is responsible for mapping resources in the sub-space to its real physical location. This P2P-style decentralized resource discovery design can help Grids ensure scalability.

The problem is that the resource searching algorithms in different P2P Grid system is diverse. Different static DHT topology designs, such as different network diameter (number of hops for a message from source to destination) or network degree (number of neighbors with which a node maintains long-term contact) often leads to different dynamic DHT algorithm (see Table I). Accordingly, it's hard to searching resources in heterogeneous P2P Grid systems. The resources search and integration problem is an important topic to P2P Grid.

DHT	Static topology	(Av.) degree	Diameter
HyperCup	Hypercubes	$O(\log N)$	$O(\log N)$
Chord	ring graphs	$O(\log N)$	$O(\log N)$
Pastry/Tapestry	Plaxton trees	$O(\log N)$	$O(\log N)$
Kademlia	ring graphs	$O(\log N)$	$O(\log N)$
Viceroy	butterfly	$O(1)$	$O(\log N)$
Cycloid	cube conn. Cycles	$O(1)$	$O(d)$, $N = d \cdot 2^d$
Koorde/D2B	de Bruijn	$O(1)$	$O(\log N)$
Distance Halving	k-base de Bruijn	$O(k)$	$O(\log N / \log k)$
CAN	d-dim. torus	$O(d)$	$O(d \cdot N^{1/d})$

Table.1 DHT and Static DHT Topologies

By considering all resources as services in a Service-Oriented Architecture (SOA) methodology, resources in heterogeneous P2P Grid systems can actualize interoperation. In such circumstance, service is the only subject that needs to be concerned about rather than the software topology or hardware components. Functional and non-functional requirements for matching capable service and concerned resources may be translated into thousands of services. SOA is facing the challenge in making meaningful and dynamic discovery, orchestration (composition), and integration (adaptation). The lack of these capabilities in SOA is called the semantic gap. Semantic technology could close this gap ideally.

Fortunately, considerable research has been completed and under way to realize the potential of semantically enabled SOA, such as OWLS [15], SWSF [16], IRS-III [17], and WSDL-S [14], each of which has gained some momentum and addressed some pragmatic aspects. Through introducing semantic technologies, this paper presents a layered search model that has a good character in semantics to solve the search problem with scalable P2P Grid.

3. Resource model

This paper proposed a four-layer resource search model called FSCP (as illustrated in Fig.1): Functional Service Layer (FSL), Semantic Service Layer (SSL), Common Service Layer (CSL) and Physical Resource Layer (PRL). Based on the FSCP resource space model, client applications can solve functional business requirements more conveniently by integrating FSL on-the-fly; the FSL indexing the multi-SSL resources for searching; SSL resources answer for merging cluster and compositing functional Web Services which can map to multi-CSL layer resources. CSL resource actually visits multi-PRL resources by web spiders. PRL is refered to the current large internet web sites. From FSL to PRL, there are some dash connections. Namely, multi-PRL resources of the same kind can be grouped into a same CSL resource. This function can help realize load balance. With FSCP, it is more effectively to describe, discover and search resources in P2P Grid.
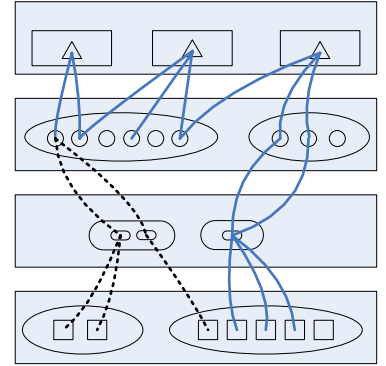

Fig.1: Resource Space Model

3.1. Physical resource layer

Definition 1 Node is an independently operated computer device that can offer service in the P2P Grid systems. It has many web resources like web pages, images and videos.

A whole resource in P2P Grids is defined by $G=(R,N,E)$. R is defined as a set of physical web resource r_i and $N = \{N_1, N_2, N_3, ..., N_n\}$ is the set of nodes. Each node N_i contains several physical web resources P_i . E is the adjacent relationship of nodes and $E(N_i)$ is the neighboring nodes of any node N_i .

Definition 2 A set of G is defined as $P = \bigcup\limits_{i=1}^{n} P_i$. Namely, the P is the set of all PRL resources.

Let p_i be any partition of $P = \{P_1, P_2, P_3, ..., P_n\}$, as to any $1 \le i, j \le n$, there is $P_i \cap P_j = \varnothing$.

Let $w = |P|$, then all physical resources G can be

defined as $P = \{p_1, p_2, p_3, ..., p_w\}$.

In actual environment, there are many resources of the same kind existed in different nodes, such as identical files. The function $f1: R \rightarrow P'$ is denoted to merge them then put forward the new resource set P'.

Definition 3 A resource r is defined as a couple (Z, D), where $Z(r)$ is defined as (N_i, r_j). $D(r)$ is the description of r.

In P2P Grid, (Z, D) is sufficient for defining resource r, and it is common for r to be in different forms of varying nodes. As to P2P Grid systems with different static topologies, an identical file will be polymorphic, so the presentation of resource r may introduce the issue of interoperation at PRL.

3.2. Common service layer

Supposing that there are m Common Services in P2P Grids: $GS_1, GS_2, GS_3, ..., GS_m$, every node N_i provides some Grid Services GS_i.

Definition 4 A Common Service ws is a quaternion (A, R, P, D) where $R(ws)$ is defined to be a universal address in CSL; $R(ws)$ is defined to be the set of Web Service parameters, $P(ws)$ represents the physical resources employed by ws; and $D(ws)$ is the description of ws.

Definition 5 Resources in CSL are $C = \bigcup_{i=1}^{m} GS_i$. Namely, the C is a set of all CSL resources.

Let C_i be any partition of $C = \{C_1, C_2, C_3, ..., C_n\}$, as to any $1 \le i, j \le n$, there is $C_i \cap C_j = \varnothing$.

For any PRL resource space $P_i = \{P_1, P_2, P_3, ..., P_n\}$, there is a 1:n refection $f2: C \rightarrow P$.

Deduction 1 Let $c = |C|$, $u_i = |GS_i|$, then there is
$c = \sum_{i=1}^{m} u_i$.

Deduction 2 For any identical physical resource $c' = f2(c) = \{P_1, P_2, P_3, ..., P_k\}$, there exists an $f3: C \times D \times D \rightarrow c'$.

Namely, for a given CSL resource c, $D(c)$ and $D(c')$, any PRL resource can be located by $P_i = f3(c, D(c), D(c))$, where $D(c')$ is the description of $c' = f2(c)$.

3.3. Semantic service layer

Supposing that there are m SSL resources $SGS_1, SGS_{12}, SGS_{13}, ..., SGS_{1m}$, where each SGS_i is grounded in some CSL resources.

Definition 6 SSL resources are defined as $S = \bigcup_{i=1}^{m} SGS_i$. Namely, the S is a set of all SSL resources.

Let S_i be any partition of $S = \{S_1, S_2, S_3..., S_n\}$, then to any $1 \le i, j \le n$, there is $S_i \cap S_j = \varnothing$.

Deduction 3 Let $u = |S|$, $u_i = |S_i|$, then
$u = \sum_{i=1}^{m} |S_i| = \sum_{i=1}^{m} u_i$, $S_i = \{s_{i1}, s_{i2}, s_{i3}..., s_{ui}\}$.

For any SSL resource space $S_i = \{s_1, s_2, s_3..., s_n\}$, there is a 1:n reflection $f4: S \rightarrow C$.

Deduction 4 Provided that any identical physical resource $c' = f2(f4(s)) = \{P_1, P_2, P_3, ..., P_k\}$, then there exists a refection $f5: S \times D \times D \rightarrow c'$.

Namely, for any given CSL resource s、 $D(s)$ and $D(c)$, any PRL resource can be located by $P_i = f5(s, D(s), D(c))$, where $D(c)$ is the description of $c = f2(f4(s))$.

3.4. Functional service layer

Definition 7 Resources in FSL are distributed indexing resources and offers search APIs.

There is a reflection, that is, one FSL resource is reflected to many semantic services.

4. Search architecture

The architecture of resource searching is discussed below based on the FSCP search model. By this kind of architecture, the resources can be organized effectively, conveniently and reliably for searching and sharing
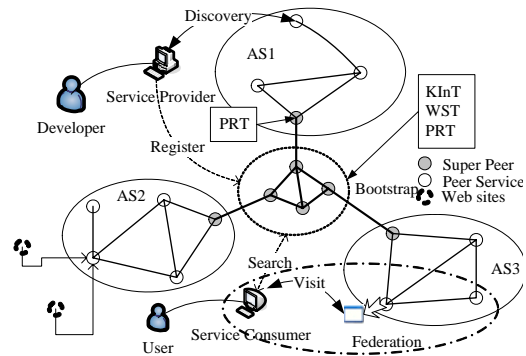

Fig. 2: Architecture of Resource Searching

4.1. User application

This kind of system involves Grid Portal, and other

Grid applications like Service Consumer illustrated in figure 2. The main function is to represent user's functional requirements then to invoke FSL resources. Such a system has a friendly user interface and instruction. When a user ask for searching a document at the Grid Portal, they do not need to understand where the resource comes from, what form the resource has and how the resource be allocated.

4.2. Bootstrap

The Bootstrap takes the advantage of Grid's stability, dependability, high-performance and high-bandwidth to establish the center information system. It is responsible for resource search in corresponding SSL and PRL. By using this system, three advantages as follows is addressed:

- Self-organization of the entire computational environment.
- The search of any SSL resource.
- Searching for PRL effectively to make sure the forwarding of queries from PRL.

The functionality of the Bootstrap is similar with a MDS of computational Grid, and it contains three tables based on DHT: Physical Resource Table (PRT), Web Service Table (WST) and Keyword Index Table (KInT).

Every record in the PRT denotes a physical resource of P2P Grid, and its data architectures are as follows:

- Node ID: It is a universal sign for each physical resource also named Service Peer ID (SPID). SPID is a string gained from the consistent hashing [5] against node's particular information (such as IP, domain name and other additional information, etc.).
- Physical Resource ID: It is a ID generated by the couple (Z, D) of a physical resource.

Web Service resource is managed by WST which has the data structure as follows:

- Web Service Resource GUID: It is a global ID generated by the consistent hashing based on Web Service's particular information, and GUID acts like a primary key in RMDB.
- Service URI: Such as https://192.168.1.23:8080/TrackerService. URI wraps the information of host name, binding protocol, service name, etc.
- WSDL-S URI: Such as https://192.168.1.2:9080/TrackerService?wsdl.

WSDL-S has proposed a more technically centered mechanism to augment WSDL by making the approach agnostic to any ontology representation language as illustrated in Figure 3. Using the extensibility elements of WSDL, a set of annotations can be created to semantically describe the inputs, outputs and operations of a Web Service.
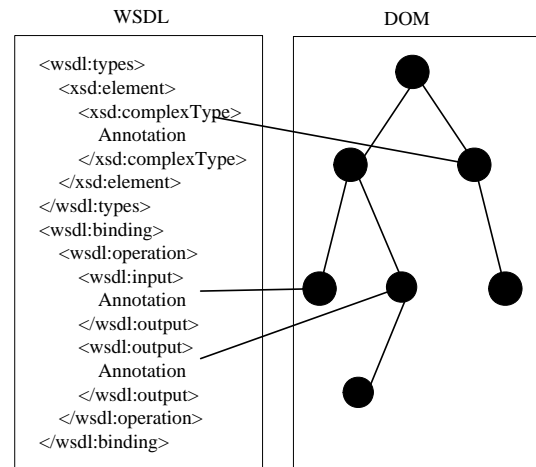

Fig. 3: Semantics WSDL Elements

There are two ways to search resources based on WST. One is to search through GUID. When the Web Service is issued, the service information (such as GUID) is injected into WST, and definitely, it is convenient and fast to use GUID to carry out the search against DHT. The other way is to search through keywords. The full-text search technology (such as suffix array, inverted indexes, signature files, etc) can be bought from the web search engine. This architecture suggests the inverted indexes. The process of using inverted indexes to carry out the search in Bootstrap through keywords is as follows: the parameter P_j of Web Service S_i contains a keyword "K", and it can compose a triple (K, S_i, P_j). All the triples which containing the keyword "K" form a list as follows: $<K, (S_{i1}, P_{j1}), (S_{i2}, P_{j2}), (S_{i3}, P_{j3}), ..., (S_{in}, P_{jn})>$ 。

The aggregation of the entire keywords list forms the index table. Dispersedly registering the index table into Bootstrap shapes so called KInT. KInT is a physical resource and the user can use this kind of resource through CSL Grid Service, which offers an RMDB-SQL like way.

Comparing to the single node in the P2P system environment, DHT in Bootstrap communicates and synchronizes through the way of Web Service, using identical static topology. DHT records in the Bootstrap cover the entire network scope and stores enduringly. Bootstrap takes advantage of its own WST to leverage service mining and assembling. WST in Bootstrap is a repository of Web Service resources. The repository is a source for assembling new services by means of combining specialized knowledge database and encapsulating it with Semantic Web.

4.3. Autonomy system

In the large-scale Internet, backbone nodes have the characteristic of power-law and Small World. Based on these features, thousands of nodes in P2P Grid could be grouped together into an Autonomy System (AS) by a certain interest or logic relationship. An AS is consists of many Gird nodes and situated in lower 2 layer of FSCP model. It maintains the lifecycle of CSL resources and Services in the AS which can be federated. AS, like KaZaa, is a hierarchical system with multi-virtual organization (VO) and each VO can also has multi-Super Peers. Strong nodes (characteristics with stable, high speed and high performance) act as Super Peer. Super Peer also maintains a physical resource table in certain range (copy of partial PRT in Bootstrap).

4.4. Service peer

The Service Peer (SP) actually exists in a physical node beyond PRL. As shown in figure 4, a SP communicates with other nodes through Peer Service and harvest web resources as a spider. Peer Service is a type of Grid Service that is located in the CSL of FSCP. By P2P Module, which is located in the CSL, SP manages its physical resource by P2P Module and interacts with other SPs through Peer Service. The SP also exposes specific functional services by Peer Service. A Client Interface can leverage users to write client applications.By Spider Module located in the CSL, SP collecting user defined web resources among the whole internet web resources.

Grid nodes exchange messages originally through socks streams over the TCP/IP. For a Peer Service is concerned, it requires nodes to interoperate through Web Service, that is, connections are also services. Peer Service must implement some given interface as follows:

1) Connect or disconnect with other nodes in heterogeneous or homogeneous P2P Grid topologies;
2) Message routing, forwarding cache and message merging in the local node before forward.
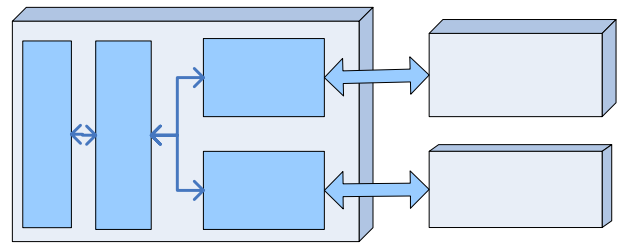3) Communicate with local P2P Module, which maintains a PRT to stores local physical resources.

Fig. 4: Structure of Service Peer

4.5. Nodes collection and routing

In the above-discussed architecture, user can integrate FSL resources into client applications. Resource merging and orchestration are conducted at SSL, key words and WSDL-S which are based on Bootstrap. Physical resources which are heterogeneous and have different static topologies are collect by spider. The actual invoking of resources will happen at PRL. The organizing protocol of physical resources will be designed upon the interaction ability of SP. The organizing protocol is message based on primary contains as follows:

Nodes Collection: It is a common event for a node to join an AS, leave an AS or adjust the level in an AS. So the nodes collection algorithm shall be effective and fast convergence. Some algorithms like SOMO [6] perform a whole system investigation to acquire statistical information to achieve the nodes convergence and emanation. The node free algorithm such as PeerWindow [8] is proposed to do nodes collection in heterogeneous environments.

Routing: By using Web Services, Peer Service takes the responsibility for routing search requests. Routing is carried out in the AS and Bootstrap through Greedy-Search (Fig.5) or other algorithms like Tourist [8] and Kademlia.

```
Greedy_Search(key)
    if key is contained in node n
        return n;
    else
        for each node  n_i  linked with n
            select a nearest  n_i
        Greedy_Search(key)
```

Fig. 5: Greedy-Search Algorithm

5. Search architecture

Based on the FSCP search model and resource management architecture above-mentioned, a prototype of vertical web search engine for e-

commence has been implemented primarily. By deploying Grid environment using GT4 with a embed Bootstrap, and exposing Web Service through AXIS, the common Grid nodes was transformed into the SP, which has the web page harvest functionality, and have designed a Grid Portal as the entrance for search. In the experiment, lots of web pages has been harvested in the real internet at different nodes.Each node harvests a part of e-commence sites from Amazon.com to Alibaba.com. Web data is collected through web crawling tool based on WebHarvest. Before the dates exchanges with other spider results through PS, date extraction is implemented using GATE[18][19]. The web resources are merged at SSL then clustered. After that, resources are indexed at FSL by Lucene.

6. Conclusions

This paper has investigated a resource search model called FSCP. With this model and its corresponding resource organizing architecture, we offer a new way for the resource searching and sharing problem existed among the large internet web sites. The Bootstrap supplies a uniform information centre to P2P Grid systems, and it enhances the resource search ability. The SSL provides a basis for realizing FGSM [9] and DSI [10]. Future work will be focused on the following three aspects: 1), optimizing the node broadcasting and physical resource harvesting algorithm; 2), study the service performance in the Bootstrap and the whole architecture. The contribution of this paper is to provide a semantic-based solution to web resource discovery, search and sharing based on P2P Grid.

References

[1] D. Booth, H. Haas, E. Newcomer, et al. *Web Services Architecture*. W3C Working Group. http://www.w3.org/TR/ws-arch/, 2004.

[2] I. Foster, C. Kesselman, J. Nick, S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. *Open Grid Service Infrastructure WG*, Global Grid Forum, June 22, 2002. http://www.globus.org/research/papers/ogsa.pdf.

[3] H. Balakrishnan, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Looking up data in p2p systems. *Communications of the ACM: Special Issue: Technical and social components of Peer-to-Peer computing*, 46(2), 2003.

[4] K Czajkowski , S Fitzgerald , I Foster et al. Grid information services for distributed resource sharing. *In Proceedings of the 10th IEEE HPDC*. Washington, DC: IEEE Computer Society Press, 2001.181～194.

[5] D. Karger, E. Lehman, T. Leighton, R. Panigrahy, M. Levine, M.,and D. Lewin, Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. *In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (1997)*, ACM Press. 654–663.

[6] Z. Zhang, S. M. Shi and J. Zhu. SOMO: self-organized metadata overlay for resource management in P2P DHT. 2nd *International Workshop on Peer-to-Peer Systems (IPTPS '03)*. 2003-11.

[7] P. Maymounkov, D. Mazieres, Kademlia: A Peer-to-peer Information System Based on the XOR Metric. http://kademlia.scs.cs.nyu.edu.2002

[8] J. F. Hu, M. Li, H. L. Yu, H. T Dong, and W. M. Zheng. PeerWindow: An Efficient, Heterogeneous, and Autonomic Node Collection Protocol. *The 2005 International Conference on Parellel Processing (ICPP-05)*. Oslo, Norway. June 2005.

[9] T. Lu, Z. S. Li, C. L. Xu, et al. FGSM: A Framework for Grid Service Mining. *Journal of Sichuan University (Engineering Science Edition)*. 2005.37(2): 86-92. (in chinese).

[10] G. W. Yang, S. M Shi, D. X. Wang, et al. DSI Distributed Service Integration for Service Grid. *J.Comput.Sci.& Technol*. 2003,18(4):474-483.

[11] I.Foster, K. Czajkowski, D.Ferguson,J.Frey,S.Graham et al. Modeling of OGSI and WSRF. *In Proceedings of the IEEE*,93(3),2005.

[12] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. *In Annual Conference of the ACM Special Interest Group on Data Communications (ACM SIGCOMM 2001)*, San Diego, CA, USA, Aug. 2001.

[13] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. *In Annual Conference of the ACM Special Interest Group on Data Communications (ACM SIGCOMM 2001)*, San Diego, CA, USA, Aug. 2001.

[14] R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M. Schmidt, A.Sheth, and K. Verma. Web Service Semantics – WSDL-S. Technical note, 2005. http://lsdis.cs.uga.edu/library/download/wsdl-s.pdf.

[15] The OWL Services Coalition. OWLS 1.1. Available from http://www.daml.org/services/owls/1.1/, November 2004.

[16] Semantic Web Services Framework. SWSF Version 1.0. Available from http://www.daml.org/services/swsf/1.0/, 2005.

[17] J. Domingue, L. Cabral, F. Hakimpour, D. Sell, and E. Motta. Irs-iii:A platform and infrastructure for creating WSMObased Semantic Web Services. *In Proceedings of the Workshop on WSMO Implementations (WIW 2004)* , Frankfurt, Germany, September 2004. CEUR

[18] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva and W. Peters. Natural Language Technology for Information Integration in Business Intelligence. *10th International Conference on Business Information Systems*, Poland, 25-27 April, 2007

[19] D. Maynard, W. Peters, Y. Li. Metrics for Evaluation of Ontology-based Information Extraction. *In Proc of WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON 2006)*, Edinburgh, Scotland.