A P2P Overlay Architecture for Personalized Resource Discovery, Access, and Sharing over the Internet

Chatree Sangpachatanaruk†
†Department of Information Science and Telecommunications
University of Pittsburgh
Pittsburgh, PA
Email: chatrees@mail.sis.pitt.edu

Taieb Znati∗†
∗Department of Computer Science
University of Pittsburgh
Pittsburgh, PA
Email: znati@cs.pitt.edu

Abstract—The next generation Internet has a potential to become a ubiquitous and pervasive medium communication of all types of information. Yet, because of its decentralized nature and weak support for semantic, the Internet remains a chaotic repository, lacking of ability to allow users to discover, extract and integrate information of interest from heterogeneous sources, and its ability to provide these users with efficient tools to manipulate and convert the discovered information into knowledge. Recently, the peer-to-peer (P2P) overlay technologies have been emerged to leverage resource discovery on the Internet by providing a scalable framework to develop efficient mechanisms to distribute, share, and access resources in a large scale, highly dynamic environments. The main objective of this paper is to leverage P2P overlay technology to enable a Personalized Web architecture for structured Internet resource access and knowledge sharing between community members of similar interests. The Personalized Web architecture uses an ontology-based representation of the resources to enable a semantic resource discovery and access that reflects the interest of the user. The basic components of the Personalized Web architecture are discussed, including the DHT-based, P2P overlay structure for indexing and resource location, the mechanisms required to support scalable routing, and the services and protocols to enable resource advertising, discovery and management.

I. INTRODUCTION

For the last decade, the Internet has been providing access to a bewilderingly large number and variety of resources, including text, audio and video files, raw scientific data, retail products, network services, and transcripts of interactive conversations. The full potential of a ubiquitous and pervasive Internet technology, however, has yet to be realized. The decentralized nature of the Internet, coupled with weak support for semantics and security, has made it difficult to enable expressive, semantically driven services and resource discovery tools, which closely reflect the user’s interest and avoid “information overload”.

Early resource discovery tools have been designed to facilitate access to information on the Internet. These tools have been used with varying degrees of success to achieve Internet resource indexing and discovery. It has become clear, however, that they fail to capture the specific needs of a given user. This is due to the fact that most of these tools can only provide uniform and equal access to all resources on the Internet. More recent tools use sophisticated algorithms to assign weights to different Web pages. These weights, however, are used merely to rank pages. As such, they do not necessarily reflect the users’ personal interests. As a result, queries issued by Internet users often produce an overwhelming number of responses, which frequently contain references to irrelevant material while leaving out more relevant ones. Furthermore, in their current form, these Internet tools provide very little support to allow the formation of knowledge communities of common interests, practices or aims.

Recently, the P2P overlay technologies been emerged to leverage resource discovery on the Internet. Typically, these technologies allow peers to deploy large-scale, self-evolving infrastructure in ad-hoc fashion, without the need for centralized management or control [1]. The ease of deployment of P2P overlay infrastructure enables the development of a new class of applications that can effectively and strategically provide services over the Internet.

The major goal of the paper is to address the shortcomings of “information overload” caused by current Internet tools and enable efficient resource discovery, service deployment and knowledge sharing among members of “communities of similar interests”. To this end, the paper proposes the concept of a Personalized Web (PW), which represents a set of a user’s personalized views of the Internet, where each view is responsive to a specific personal need or interest of the user.

Conceptually, a view represents a Web of Internet resources associated with a specific ontological content that reflects a specific user interest. The mechanisms used to create a PW view revolve around the concept of active advertising and discovery, whereby a semantic-aware agent, referred to as Semlet, acts as a proxy for a specific resource and uses the metadata associated with the resource, along with a user-defined similarity metric, to identify and join communities of similar interest. The discovery of new resources, coupled with the development of new views, allow a PW to evolve...
dynamically to include other PWs, thereby enabling “natural” formation of user communities of similar interests.

The focus of this paper is on developing an overlay-based architecture of a Personalized Web (PW), and on designing the required protocols and mechanisms to support the required functionalities of a PW. The remainder of this paper is organized as follows: Section II describes the PW framework and functionalities. Section III discusses major components of a PW architecture. Section IV presents main algorithms and protocols to support the functionalities of a PW. Section V discusses the related works. Finally, Section VI concludes by describing the future work.

II. PW FRAMEWORK AND FUNCTIONALITIES

In the Personalized Web framework, each resource is described by a metadata, which can be viewed as a user-defined “personalized profile” of the resource. The metadata contains at least one semantic attribute which can be mapped to a semantic reference in the ontology network. This attribute can be expressed in a simple form, using keywords for example, or it can also be described in terms of an elaborate RDF schema, which in turn is translated into an ontology.

To enable a PW, efficient protocols and mechanisms must be developed to:

i) Enable resource discovery, based on user-specific interest profiles, in a user-transparent manner,

ii) Enable formation of “communities of similar interests”, in a dynamic and user-transparent manner,

iii) Enable automatic advertising of users’ resources to “communities of similar interests”, and

iv) Manage the user’s PW to dynamically capture evolving user’s interests as the environment changes.

To support the above functionalities, we propose a PW architecture, which leverages three emerging technologies, namely P2P overlay networks, agent technologies, and semantic web. P2P overlay technology offers the capabilities to organize Internet resources into a collection of logical overlay networks. Each overlay network can be viewed as a community of peers whose members share common interests, practices or aims. In the following, a detailed discussion of the design decisions and basic components of the proposed architecture is provided.

III. PERSONALIZED WEB ARCHITECTURE

The basic design goal of the PW architecture is to provide users with the ability to adaptively create their own views of the Internet to meet their personal interest in an effective and scalable manner. To achieve this goal, the proposed approach augments the functionalities of a structured overlay network architecture with the capabilities of an agent-based technology to provide a flexible and scalable architecture capable of supporting efficiently the main PW design requirements. The basic architecture is depicted in Figure 1. The architecture comprises three logical layers, namely PW Agent layer, Ontology Overlay Network (OON) layer, and the Internet Resource (IR) layer.

The first two layers rely on the IR to physically locate and access resources. The IR layer uses currently available Internet capabilities to provide these functionalities. In the following, we discuss in more details the PW Agent layer and the OON layer.

A. PW Agent layer

The main goal of a PW is to enable a user-defined Internet structure for resource discovery and access. The PW must reflect as closely as possible the user’s interest and allow the underlying structure to evolve dynamically and in a transparent manner from the user. To achieve this goal, the proposed architecture uses an agent-based approach to support the functionalities required to allow users to express personalized views of the Internet, advertise their resources to the communities of similar interests, and discover new resources of interest based on their personalized similarity profiles.

The basic tenet of the PW agent layer is a semantic-aware agent, referred to as a “Semlet”, which represents a semantic outlet for a user resource. The Semlet acts as a “surrogate” agent for the associated resource, and as such, enables this resource to discover and advertise itself to semantically similar resources. A scope can be used to limit the actions of a Semlet to specific domains of interest, such as an Internet domain, a user community or an interest group. The Semlet uses the user’s interest profile and a user-defined measure of “semantic similarity”, to identify and join communities of similar interest within the user defined scope.

Formally, a Semlet can be defined as follows:

Semlet = (res, profile, scope, policy, ontology)

where:

• res uniquely identifies the resource to be located or advertised,
• profile defines a set of conceptual relations from a concept space characterizing the resource,
• scope delimits the “boundaries” within which the Semlet can advertise and from which the Semlet can recruit resources of interest,
• policy is the advertising strategy, which can be either aggressive, whereby the Semlet proceeds to advertise its
associated resource immediately, or passive, whereby the
Semlet waits to be discovered by other Semlets, and
• ontology identifies the ontology supported by the PW
architecture.

Semlets use the user-described profile to extract and dynam-
ically build metadata reflecting the concepts associated with
resources of interest to the user. The metadata is then used
to advertise the resource to users of communities of interest,
or locate resources of interest to the user. The capability of a
Semlet to join other groups of Semlets with similar interests
provides the basis for the development of a dynamic and
distributed graph of semantically related resources that can
be retrieved on demand in response to a user’s queries.

B. Ontology Overlay Network Layer

The PW architecture must allow for the user’s personal
views of the Internet to evolve dynamically as new resources
of interest to the user are created and advertised. Furthermore,
this process must occur transparently from the user, while
ensuring that these views continue to reflect the user’s interest
reliably. To achieve this goal, the proposed architecture uses
an ontology overlay network (OON), as a basis for resource
representation, and to enable the development of efficient
resource indexing and discovery algorithms. In this context,
an ontology is defined as a formal, explicit specification of
a shared conceptualization, which can refer to the shared
understanding of some domains of interests [2].

The underlying OON architecture is a structured, Dis-
tributed Hash Table (DHT) based P2P overlay infrastructure
that integrates Wordnet ontology tree (WOT) into its struc-
ture [3]. A WOT is represented by a hierarchy tree, where
a node in the tree represents a noun concept described by
a set of similar words (synset), and where an edge between
two nodes represents the is-a relation between these concepts.
Partial WOT is shown in Figure 2. It is used to ensure
global consistency in semantic classification. As such, given a
keyword, the architecture can consistently identify the related
concepts. Each node in WOT is assigned a numerical ID that
is used as the hash input for a search in ontology overlay
network.

The DHT is used to infer a structure that regulates location
and access to a distributed set of resources. Consequently,
the OON supports the capability to enable the development of
adaptive and scalable search and retrieval algorithms. Using
the DHT information, overlay nodes, which host a resource
of interest, can be identified and located deterministically and
in a finite number of steps, thereby reducing considerably the
search overhead due to communication and routing.

In the OON, the resource metadata is clustered, indexed
and distributed among overlay nodes based on the semantic
described by associated ontology. Note that only the metadata
associated with the resource is distributed among nodes in the
OON. The actual resource remains physically located at its
original site.

IV. PW ALGORITHMS AND PROTOCOLS

In this Section, the mechanisms and data structures used
to support main functionalities of a PW will be described.
First, the PW search scheme is introduced. Following, the
mechanism and data structures used to form ontology overlay
network will be discussed. Finally, the mechanisms and data
structures to support PW resource advertising and discovery
will be explained.

A. PW Search Scheme

PW architecture is an agent-based system which achieves
its semantic richness through the use of explicit ontologies
to represent resources. PW architecture further enhances the
DHT-based resource distribution scheme by using the unique
identifier assigned to each ontology as a key to locate the
overlay node responsible for maintaining the resource index
associated with the underlying ontology. In other words,
the ontology-based hashing scheme, thereof referred to as
“Semantic-Driven Hashing (SDH)”, utilizes ontologies, in-
stead of resource names, as the hash input to generate the key
necessary to distribute the resource among overlay nodes.

When a query to locate or advertise a resource is issued,
the SDH scheme consults the WOT to obtain the ID of the
synset of the concept associated with the resource specified
in the query. This ID is then hashed to obtain the key, in the
DHT space, of the overlay node where the metadata resource
index is maintained. This key is then used to route the request
to the identified overlay node.

To further explain the SDH-based resource discovery
scheme, we consider an ontology that describes the concept
of “telecommunication”, as depicted in Figure 3. Based on
the above ontology, a given Internet user can express personal
interest in communication by using the keyword “telecommunication”. When this keyword is issued as part of a search, the resulting query returns information of all resources which are directly related to the associated concept, namely telephone, telegraphy, wireless, broadcasting, and multiplex. When used in conjunction with the keyword of “wireless”, however, the ontology can be scoped down to the resources that are associated with the new concepts (i.e. telecommunication that is wireless), thereby capturing more specific the user’s interests.

B. OON Setup and Organization

This section describes the procedure used to set up an OON. First, the system model is described. The organization of the OON into a SDH-based overlay structure is then discussed.

To support semantic-based resource indexing, discovery and advertising, the OON structure assumes the model:

- The Internet is organized into a set, \( D \), of autonomous domains. Each domain, \( d \in D \), elects a node, \( n_d \), to act as a core node in the semantic overlay network. The set of core nodes, referred to as \( N \), forms the basis of the OON.
- In the OON structure, each node, \( n_d \in N \), has a set of neighbors, \( N_d \), which it uses to route data using a DHT-based strategy.
- Nodes in a given domain, \( d \), address their search and advertising queries to core node \( n_d \in N \).
- Let \( R \) be the set of resources in \( D \) and \( K \) be the set of all keywords. Each resource \( r \in R \) is represented semantically by a set of keywords \( K_r = \{ k_i^r \in K, 1 \leq i \leq m \} \). Furthermore, each resource \( r \in R \) is associated with a set of metadata which contains profile information, such as the URL and other attributes specific to the type of the resource.
- Let \( C \) represent the set of concepts in WordNet. A WOT synset, \( S_i^c \), relative to a concept \( c \in C \), is defined as a set of keywords, \( S_i^c = \{ k_i^c \in K, 1 \leq i \leq m \} \), such that \( k_i^c \) maps to concept \( c \). We use \( S \) to denote the set of all synsets in WordNet.
- A keyword, \( k_i \), may refer to multiple WordNet concepts.

Each core node \( n \in N \) is assigned a set of concepts and has the responsibility to store the metadata associated with the resource. The core node uses the metadata to locate the resource associated with the concepts it stores, and support the management of the communities of similar interests. We use COMM to denote the data structure containing relevant information to a community of interest, such as the creator and its profile information, and a pointer to next member of the community. The concept mapping to core nodes is achieved using procedure, Concept_Resource_Map(), described as follows:

Procedure Concept_Resource_Map()
  - For a given resource, \( r \in R \), and for each keyword \( k_i^r \in K_r \), use the mapping \( WN() : K \rightarrow C \) to determine the set of concepts \( C_i^r \) associated with \( k_i^r \).

Let \( C_r = \bigcup C_i^r, 1 \leq i \leq ||K_r|| \), be the set of concepts associated with resource \( r \).

- Concept Mapping into Core Nodes.
  - For a given concept, \( c \in C \), use the mapping \( SDH() : C \rightarrow N \) to determine \( N_r = \{ n \in N, \) such as \( n \) is responsible for managing the resources associated with concept \( c \} \).
  - Each node \( n \in N_r \) stores the metadata associated with resource \( r \).

The following section describes the procedure used to advertise and discover resources. A description of the basic steps used for community of interest creation and information dissemination among members of a community.

C. PW Resource Discovery and Advertising

A PW evolves based on the interaction of the Semlets in communities with similar interests. Each Semlet manages resources of interest in the context of the associated user profile. A user’s Semlet supports three main functionalities. First, the Semlet is responsible for advertising the user’s resources to communities of similar interest, and discovering new resources as they are made available by these communities. The Semlet uses the metadata associated with the resource to carry out this responsibility. Second, the Semlet may create communities of similar interests, as new interests of the user arise. Finally, the Semlet must cooperate with other Semlets to allow the user’s PW to evolve effectively and in a transparent manner from the user. A detailed discussion of the main design issues to support these functionalities is provided next.

1) Semlet Advertising: The main objective of Semlet advertising is to identify all users who share interests similar to the ones associated with the resource. The Semlet can then advertise its profile to these users, thereby avoiding information overload. The following describes the basic steps required to achieve this goal.

- Profile Development:
  - Let \( R_u \), be the set of resources owned by user \( u \). Using \( R_u \), a profile, \( P_u \), is obtained as a superset \( K_u \) of all keyword sets associated with each resource \( r \in R_u \). \( P_u \) can be defined as follows:
    \[ P_u \leftrightarrow K_u = \{ K_r, r_i \in R_u, 1 \leq i \leq ||R_u|| \} \]

- Synset Development:
  - Let \( C_u \) be the concepts of interest associated with user \( u \). Use the mapping \( WN() : K \rightarrow C \) to compute of concepts \( C_i^u \) associated with \( k_i^u \in K_u \). Compute \( C_u = \bigcup C_i^u, 1 \leq i \leq ||K_u|| \).

In order to advertise and retrieve resources of interest, the Semlet identifies the OON nodes responsible for managing metadata of all resources associated with each concept \( c \in C \). The Semlet queries these OON nodes to obtain the metadata of all these resources. Using this metadata, the Semlet applies the user defined similarity metric to compute similarity score for each resource. These scores are then used by the Semlet to determine communities of similar interests. The Semlet uses
the set of identified communities to advertise its resource to their members, and retrieve metadata of interest to the user.

The similarity metric used by the Semlet to identify resources of interest verifies two main properties, namely asymmetry and non-transitivity.

- Asymmetry is due to the fact that each Semlet has its own ontology and the user defined metric to identify resources of interest. Consequently, the interest of Semlet A in resource B does not imply that Semlet B has interest in resource A.
- Non-transitivity also stems from the fact that each Semlet uses its own ontology and similarity metric. Consequently, the interest of Semlet A in resource B and the interest of Semlet B in resource C does not necessarily imply that Semlet A has interest in resource C.

Mindful of the above properties, Semlets independently develop their communities of interests. For example, Semlet A and B have resource profiles \( r_A \) and \( r_B \). Semlet A is interested in resource \( r_B \), when \( Sim^A(r_A, r_B) \leq \delta_A \), where \( Sim^A(\cdot) \) is the similarity function for Semlet A and \( \delta_A \) is a similarity threshold for Semlet A. On the other hand, Semlet B is interested in resource \( r_B \), only if \( Sim^B(r_B, r_A) \leq \delta_B \).

The similarity threshold (\( \delta \)) is used to trigger advertising to other Semlets. To efficiently manage its groups of interest, each Semlet forms an advertising multicast tree MT for the members who have interest in its resource. The algorithm used to build a MT is described below.

Build MT for resource \( r_u \):

\[
\text{for all } c_k \in C_r \text{ do} \\
\quad SDH(c_k) \rightarrow n \in \mathbb{N} \\
\quad L(c_k, n) \text{ retrieve leaders } COMM_{c_k} \text{ from node } n \\
\quad \text{for all member } m \in COMM_{c_k} \text{ do} \\
\quad \quad \text{if } Sim^m(r_{c_k}, r_u) \leq \delta_u \text{ then} \\
\quad \quad \quad \text{add } m \text{ to } MT^m_{c_k} \\
\quad \quad \text{end if} \\
\quad \quad \text{if } Sim^m(r_{c_k}, r_u) \leq \delta_m \text{ then} \\
\quad \quad \quad \text{add } r_u \text{ to } MT^m_{c_k} \\
\quad \quad \text{end if} \\
\quad \text{end for} \\
\text{end for}
\]

2) Community of Similar Interest: After the overlay nodes that are responsible for related concepts are located, the Semlet visits each overlay node to retrieve all community profiles. If a community associated with a concept of interest does not currently exist, the Semlet sets up the structure to create this new community. The Semlet is then designated as the owner of the community.

\[
\forall c_k \in C_o \\
\text{Find } L(c_k) = \{l_1, ..., l_j / l_1 \text{ is the primary leader of } COMM_{c_k}, l_2, l_3, ..., l_j \text{ are secondary leaders of } COMM_{c_k}\} \\
C_i \Rightarrow \{ \exists COMM \text{ return addresses of the COMM leaders} \}
\]

Each community is formed and managed by a group of Semlets that have similar interests. However, because of the asymmetric and non-transitive properties of Semlets, members in a community may have different views of resources of interest. In the following, the adjustments of multicast tree and community of similar interest in three major scenarios will be described.

- If \( u \) selects a resource from any members of the \( COMM \), and the \( COMM \) members select a resource from \( u \), \( \rightarrow u \) becomes member of the \( COMM \).

\[
\exists r_k^m \in MT^m_u \text{ and } \exists MT^j_{k_i} \ni r_k^u \ni r_k^i \in COMM^i_{k_i}
\]

- If \( u \) selects a resource from any members of \( COMM \), and no \( COMM \) member selects a resource from \( u \), \( \rightarrow u \) creates a new \( COMM \) and requests \( n_i \) to setup a new \( COMM \) with itself as the leader.

\[
\exists r_k^m \in MT^m_u \text{ and } \not\exists MT^l_{k_i} \ni r_k^u \ni u \in COMM^l_{k_i}
\]

3) PW Management: Based on a user’s interest profile, Semlets dynamically and independently manage a collection of MTs for the user. All the MTs of a user will then develop a PW to closely reflect the user’s interest. An example of a PW developed from a user profile \( P_u \) is shown in Figure 4.

![Fig. 4. A Personalized Web for user u](image)

V. RELATED WORKS

Both structured and unstructured overlay network architectures have been proposed for resource sharing and load balancing in a peer-to-peer environment. Search mechanisms based on unstructured P2P architectures range from simple global flooding, blind search, to complex selective flooding. Despite its simplicity, blind search can be costly in term of network bandwidth. Furthermore, there is no guarantee that existing resources will be located. A supernode-based, hierarchical P2P architecture introduces the concept of search hubs in an effort to reduce the amount of bandwidth required for flooding and to cluster indexes of all shared resources of all nodes in the same group [4], [5]. This approach achieves better scalability, but at a higher cost of supernet management. Furthermore, the scheme, which still uses blind search to locate resources, may
lead to inconsistent outcome, depending on the search starting point. The recent trend to improve resource discovery for the unstructured P2P networks is to make semantic information from all nodes recognizable and form a semantic network used as a guiding multicast. Information resulting from a search or query is cached and used to form a semantic overlay network [6], [7], [8], [9]. Other frameworks combine ontology-based data representation with P2P technology to develop an overlay, semantic network [10], [11], [8], [12]. Like PW architecture, these frameworks use a semantic overlay network to optimize resource discovery. PW architecture, however, uses a unique DHT-based, underlying P2P structure to support deterministic data location and routing. This is achieved by striking a balance between the cost of deterministic data location and the complexity of managing semantic-based resource indexing and retrieval. As result, PW architecture provides effective, semantically-based mechanisms to deterministically locate resources, without incurring substantial overhead.

On the other hand, the recent trend to improve resource discovery for the structured P2P networks is to enable semantic-based data location and search by associating semantics into the structure to regulate resource distribution. Recent frameworks propose to use information retrieval techniques, namely vector space model (VSM), and latent semantic indexing (LSI), to obtain semantic representation from the textual information appearing in the resource and queries [13]. Another approach uses XML techniques to generate semantic representation from metadata attributes [14]. The most similar work to the SDH proposed scheme is the approach that uses a concept of global catalog to provide a standard for semantic classification [15]. While the first approach does not work well with non-textual resources, the second approach is likely to provide only perfect-matched result for a search, because of the indexing and searches that depend on keyword attributes. The catalog approach even though provides a better scheme for semantic-based resource classification, its naming still depends on the keywords. The PW architecture, on the other hand, uses global classification, but contrary to the catalog-based approach, the classification does not depend on word appearances. Instead, it relies on the ontologies that capture the keywords’ meanings and semantic relations, which can be used to infer related resources more intuitively.

VI. CONCLUSION AND FUTURE WORKS

This paper focuses on Personalized Web architecture, which leverages P2P overlay network to enable personalized resource discovery, and access Internet resources to reflect users’ interests. The main components of the Personalized Web architecture are Wordnet ontology, DHT P2P overlay infrastructure “OON”, and a personalized agent “Semlet”. The Wordnet ontology provides semantic representation necessary for resource discovery that reflects the user’s interest. It is used in OON to regulate the distribution of resources and communities in the DHT overlay. As such, the resources of interest and their associated user communities are located effectively. A Semlet, on the other hand, interacts with the OON and other Semlets to dynamically develop a web of resources of interest for the user.

Over all, this paper provided concepts and high level description of scalable search and community formation mechanisms of Personalized Web architecture. Main research issues remain how to develop the architecture that is scale with high load, fault-tolerance, and trustworthy. One of the main concerns is how to deal with overloaded key in DHT network, since the PW architecture partitions network based on concepts, whose popularity among them are not uniformly distributed. Caching can help to serve search queries in some extent. However, storage load balancing strategy is needed. In addition, other issues, for example, how to recover data when the node that responsible for a popular key leave the network, and how to provide protection to the data of metadata index and user communities in the overlay, need to be addressed. Furthermore, the DHT network operations such as bootstrapping, node join and leave, need to be revisited to design a seamlessly implementation.

REFERENCES