Advanced Engineering Forum *Online: 2011-09-09*
ISSN: 2234-991X, Vol. 1, pp 305-309
doi:10.4028/www.scientific.net/AEF.1.305

Indexing Associated Knowledge Flow on the Web*

Xue Chen[1,2,3,a], Fang Tao[4], Wu Chao[1]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai

[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing

[3]State Key Laboratory of Software Engineering, Wuhan University, Wuhan

[4]School of Adult Education, Henan University of Economics and Law, Zhengzhou

[a]xuechen@shu.edu.cn

Keywords: Associated Knowledge Flow, P2P, Web.

Abstract. The Associated Knowledge Flow (AKF) on the Web is an ordered sequence of Web pages that have associated relation. The associated relation from page A to page B indicates that users who have browsed page A is likely to also browse page B. The motivation of this paper is to index the AKFs on the Web and provide users AKFs instead of discrete resources. We build a scalable P2P-based Web resource-sharing system and design two kinds of ID spaces (hash ID space and semantic ID space) on it to index resources and facilitate AKF discovery. Theoretical analysis and simulations show that such a system can achieve logarithmic performance and cost.

Introduction

Current search engines such as Google and Yahoo! mainly offer keyword-based search service and the returned answers are a group of discrete Web pages. In fact, these Web pages may have some semantic relations with each other, so connecting semantically-associated pages into the form of flows may not only facilitate users' browsing and understanding, but also help users acquire more precise answers [1]

The motivation of this paper is to provide users Associated Knowledge Flow (AKF), an ordered sequence of Web pages that have associated relation. The associated relation from page A to page B is in the form $A \xrightarrow{\omega} B$, indicating that users who have browsed page A is likely to also browse page B with probability ω [2].

Prof. Luo et al. develop a method to discover the associated relation between texts in a text set of the same domain [2]. Each text is denoted as an E-FCM (Element Fuzzy Cognitive Map) to express keywords and the associated relations between keywords. This paper employs Luo's method to extract the associated relations between Web resources. Resources can be unstructured, semi-structured or structured as long as their descriptions, tags or annotations are available to build E-FCM.

Web pages together with their associated relations constitute the Association Link Network (ALN) on the Web, which is a Semantic Link Network (SLN) built by mining the associated relation between Web pages [2][3]. The distinct phenomenon of ALN is that its degree distribution is very unbalanced. A few Web pages have a tremendous number of links (associated relations) to others, whereas most Web pages have just a few, which makes ALN hard to reach a balanced index. Our previous work has proved that the degree distribution of ALN follows a power-law form [3].

The basic idea of providing AKF service is to build an associated overlay upon the Web to index AKF. The key issues are threefold: The first is how to design the overlay topology. The second is how to organize and manage resources on such an overlay. The third is how to discover the AKFs on such an overlay.

We have proposed a ring-structured P2P topology HRing based on the Harmonic Series [4]. This paper proposes a two-layered HRing structure as the associated overlay to index AKFs. The domain names of resources are prefixed to their IDs so that resources of the same domain can be organized in neighboring HRing nodes. Two hash functions, consistent hash (CH) and locality sensitive hash

(LSH), are used to generate the two kinds of suffixes of resource IDs. CH can make resources uniformly decentralized to balance load among HRing nodes. LSH can make semantically-close resources of the same domain organized in the same nodes or in the neighboring nodes [5]. Theoretical analysis shows that discovering AKFs on HRing can achieve logarithmic routing table size and routing hops.

Related Work

The construction methods of P2P topology can be roughly categorized into four types: DHT (Distributed Hash Table) topology, tree based topology, small-world based topology and SkipList based topology.

DHT topology can uniformly map P2P nodes and resources into a single ID space, and make each nodes manage a set of resources whose IDs belongs to a specific range [4]. Balanced binary tree topology can be used to improve search efficiency by building vertical and horizontal links in each node, thus preserving resource semantics and locality [6]. Skip-List-based overlays such as SkipNet and Skip Graph support range query[7][8][9]. They can achieve $O(\log(n))$ routing hops in expectation with $O(\log(n))$ routing table size. Small-world based topology adds long links based on the distance between nodes following a harmonic probabilistic distribution, which can reach logarithmic routing hops but requires global information on the network size [10].

HRing topology is a small-world based one that HRing can achieve both high performance and low maintenance cost, while guaranteeing remarkable robustness. And more importantly, the construction of HRing topology is entirely independent of the ID space, thus independent of the upper applications. It supports coexistence of multiple ID spaces. Thus, HRing can serve as the associated overlay on the Web to discover AKFs among decentralized heterogeneous resources.

The Architecture of the Associated Overlay

Nodes are designated to download resources of specific domains on the Web. Each domain captures a big content category such as movie, music and sports, etc. Nodes on the associated overlay are identified by node IDs. Fig. 1 illustrates the node ID structure, which composes a 32-bit domain ID and a 32-bit hash ID. Both use the consistent hash function SHA-1. The domain ID is obtained by hashing the domain name, and the hash ID is obtained by hashing the node's IP address. So nodes can be linearized into a ring structure in order of their 64-bit IDs.

Fig. 1. Node ID structure on associated overlay

Fig. 2 shows a two-layered associated HRing overlay. The first layer is the whole HRing, where resources of three domains are managed on red nodes, blue nodes and green nodes respectively. Hashing domain names as the prefixes of node IDs guarantees that nodes that store resources of the same domains are neighboring each other. The second layer is three sub-HRing managing resources of three domains. Each node has two routing tables corresponding to the two layers. Fig. 2 illustrate the two routing tables of node ID10. Since the construction of HRing topology is independent of the applications, so the routing table construction on HRing does not rely on node ID space and resource ID space. Due to space constraints, we here cannot elaborate the routing table construction process. Please see [3] for more detail.

Resource ID Management on HRing

A resource is denoted as an E-FCM, a collection of keywords and the associated relations between keywords (see Fig. 3). Each resource has two IDs: hash ID and semantic ID. As shown in Fig. 4, hash ID is composed of 32-bit domain ID and 32-bit CH ID, while semantic ID is composed of 32-bit

domain ID and 32-bit LSH ID. The prefix domain ID is obtained by SHA-1 to hash the domain names. The CH ID is obtained by SHA-1 to hash keyword sequences of E-FCM, aiming to distribute resources uniformly to keep load balance among HRing nodes. The LSH ID is obtained by LSH to hash keywords and relations of E-FCM, ensuring that semantically-close resources are clustered in the same nodes with high probability.

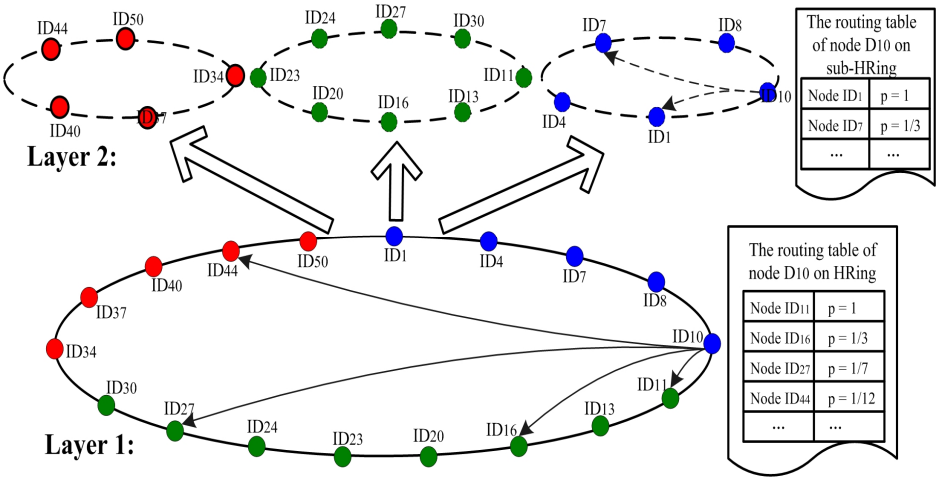

Fig. 2. A two-layered associated HRing network

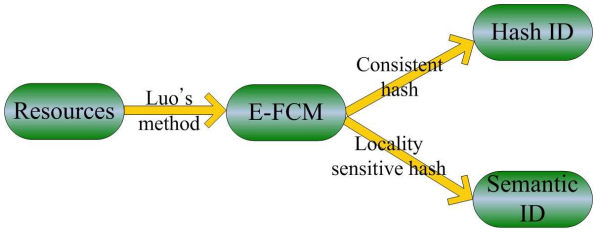

Fig. 3. ID conversion process

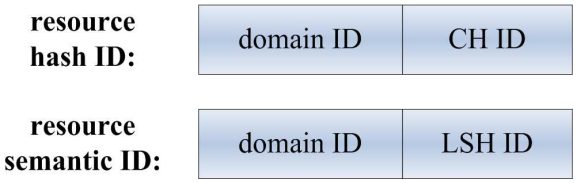

Fig. 4. Resource ID structure

The LSH ID is obtained as follows. Suppose the E-FCM of a resource A is a set of keywords and their relations $A=\{a_1, a_2, ...a_n, a_i \to a_j, ... a_k \to a_m\}$. First, a binary set $A'=\{a_1', a_2', ...a_n', a_{ij}', ... a_{km}'\}$ is obtained by consistently hashing each element of A, where $a_i \to a_j$ is hashed as a string $a_i a_j$ Then the LSH ID is obtained by randomly choose one element of A'. Thus, for two resources A and B, the probability that A and B are clustered to one node is

$$p = Sim(A, B) = Sim(A', B') = \frac{|A' \cap B'|}{|A' \cup B'|}$$

which obeys the Jaccard set similarity measure.

On HRing, resources are organized according to their hash IDs (see Fig. 5). Additionally, to facilitate AKF discovery, the hash ID of a certain resource A also index the associated relation pairs where A may be the causal key and the effect key. For example, (hash ID11, ω11) shows that the resource B whose hash ID is ID11 has the associated relation with A with weight ω11, i.e., $B \xrightarrow{\omega_{11}} A$. Semantic IDs are the index of resources, which manage the hash IDs whose corresponding resources have the same semantic IDs by LSH. Thus, the prefix domain ID for node IDs and resource IDs is designed to ensure that the resource locations and the resource indexes are within the same sub-HRing.

Search Process on HRing

Users' input is allowed to be in various forms such as keywords, a paragraph of description, or even a document as long as they can be expressed as an E-FCM by Luo's method. As illustrated in Fig. 6, the E-FCM is then converted into the corresponding semantic ID using LSH. Through the

semantic ID, we can find a group of hash IDs of semantically close resources on the layer 1 of HRing. The domain prefix of node IDs and resource IDs guarantees that these semantically close resources necessarily belong to a single sub-HRing of the layer 2. Thus, we can route on a sub-HRing by hash IDs for the resources and their relation pairs. The discovery processes of the relation pairs are still within this sub-HRing, and can be stops when the relation weights are lower or the length of AKF is longer than a predefine threshold.

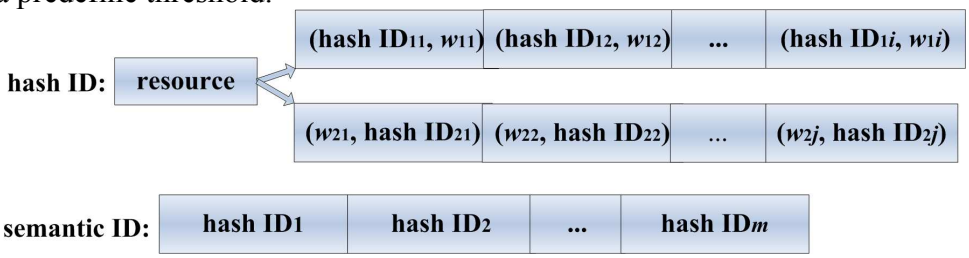

Fig. 5. Organization of resources and their index

Fig. 6. Users' input processing process

Theroem 1. On an n-size HRing overlay that manage resources of m domains, the expected routing table size is $(2\ln(n) - \ln(m))$, and the expected routing hops for discovering s AKFs of l length is $(1+sl)\ln(n) - sl\ln(m)$.

Proof. As illustrated in Fig. 2, nodes on HRing have two routing tables. Suppose that resources of each domain are managed on almost the same number of nodes. So the size of sub-HRing on layer 2 is (n/m). According to the Theorem 1 in [3], the expected routing table size on layer 1 is $\ln(n)$, and on layer 2 is $\ln(n/m)$. So the total routing table size is $(2\ln(n) - \ln(m))$. Routing by a semantic ID on HRing (layer 1) will obtain s hash IDs, which costs $\ln(n)$ hops. Then routing on a sub-HRing (layer 2) l times for each of s AKFs costs $sl\ln(n/m)$ hops according to the Theorem 1 in [3]. Thus, the expected total routing hops is $(1+sl)\ln(n) - sl\ln(m)$.

We build a 100-node HRing network as a simulation platform to index and search AKF. Each file is represented by an E-FCM of 5 keywords, which are randomly selected from 1,000 keywords. In this way, 1,000,000 files are generated since the majority of files share keywords for their associated relations. Then files are mapped into its corresponding HRing nodes using 128-bit MD5 hash function and LSH. Although MD5 is consistent and uniform, Fig.7 shows that the number of files indexed by each node is not balanced. This phenomenon is acceptable since files that have associated relation are mapped into the same nodes, in other words, several files may be given a single ID (see Fig.8). Since users' input is randomly selected keywords from the above keyword set, the most similarity between users' inputs and search outputs are below 0.5(see Table 1).

Summary and Future Work

This paper built an associated overlay based on HRing topology to provide AKF service on the Web. Two ID spaces (hash ID space and semantic ID space) are designed to identify resources on HRing. Hash ID is used to locate specific resources as well as their relations, while semantic ID is used to index a group of semantically-close resources. Theoretical analysis shows that such a system can achieve logarithmic routing table size and routing hops. This paper is the first step to organize Web resources to facilitate AKFs discovery. A lot of interesting problems such as index consistency, maintenance and load balance strategies need studying in our future work.

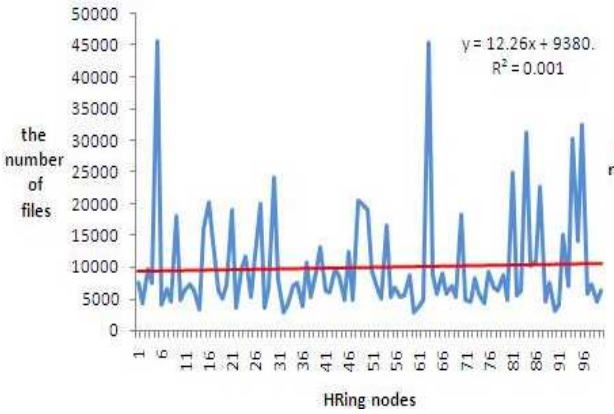

Fig.7 The load for each node

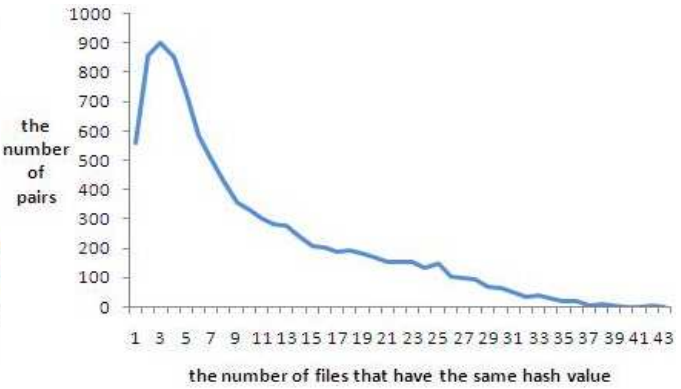

Fig.8 The number of files that corresponds the same IDs

Table 1. The similarity distribution for similarity search

query \ similarity	(0, 0.2]	(0.2, 0.5]	(0.5, 1]
10	3	6	1
50	25	24	1
100	51	38	11
200	92	88	20
300	137	133	30
500	242	212	46

This work is supported by the National Natural Science Foundation (grant number: 61001163) and State Key Laboratory of Software Engineering (SKLSE).

References

[1] X. D. Song, Y. Chi, K. Hino, B.L.Tseng, Information flow modeling based on diffusion rate for prediction and ranking, World Wide Web Conf., pp. 191-200, 2008.
[2] X. F. Luo, Z. Xu, J. Yu, F. F. Liu, Discovery of associated topics for the intelligent browsing, IEEE Conf. on Ubi-Media Computing, pp.119-125, 2008.
[3] X. Chen, X.F. Luo, S.X. Zhang and Z. Xu, Analysis and Modeling of the Semantically Associated Network on the Web, Concurrency and Computation: Practice and Experience, vol.22, no.7, pp.767-787, 2010.
[4] H. Zhuge, X Chen, X. P. Sun, E.L.Yao, HRing: a structured P2P overlay based on Harmonic Series, IEEE TPDS, vol.19, no.2, pp.145-158, 2008.
[5] Y.W. Zhu, H.H. Wang, Y.M. Hu, Integrating semantics-based access mechanism with P2P file systems, IEEE P2P Computing, pp.118-125, 2003.
[6] H. V. Jagadish, B. C. Ooi and Q. H. Vu, BATON: A Balanced Tree Structure for Peer-to-Peer Networks, Proc. Proc. 31th Int'l Conf. on Very Large Data Bases (VLDB), pp. 661-672, 2005.
[7] W. Pugh, Skip Lists: A Probabilistic Alternative to Balanced Trees, Communications of the ACM, vol.33, no.6, pp.668-676, 1990.
[8] N. J. A. Harvey, M. B. Jones, S. Saroiu, M. Theimer and A. Wolman, SkipNet: a Scalable Overlay Network with Practical Locality Properties, Proc. 4th USENIX Symposium on Internet Technologies and Systems (USITS), pp.113-126, 2003.
[9] J. Aspnes and G. Shah, Skip Graphs, Proc.14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 384-393, 2003.
[10] J. Kleinberg, the Small-World Phenomenon: an Algorithmic Perspective, Proc. 32nd ACM Symposium on Theory of Computing (STOC), pp. 163-170, 2000.