

Multimedia Correlation Analysis in Unstructured Peer-to-Peer Networks*

Bo Yang⁺ Ali R. Hurson⁺

⁺*Dept of Computer Science and Engineering
Pennsylvania State University, PA 16802
{byang,hurson}@cse.psu.edu*

Yu Jiao[§] Thomas E. Potok[§]

[§]*Oak Ridge National Laboratory
Oak Ridge, TN 37831
{jiao, potokte}@ornl.gov*

Abstract

Recent years saw the rapid development of peer-to-peer (P2P) networks in a great variety of applications. However, similarity-based k-nearest-neighbor retrieval (k-NN) is still a challenging task in P2P networks due to the multiple constraints such as the dynamic topologies and the unpredictable data updates. Caching is an attractive solution that reduces network traffic and hence could remedy the technological constraints of P2P networks. However, traditional caching techniques have some major shortcomings that make them unsuitable for similarity search, such as the lack of semantic locality representation and the rigidity of exact matching on data objects. To facilitate the efficient similarity search, we propose semantic-aware caching scheme (SAC) in this paper. The proposed scheme is hierarchy-free, fully dynamic, non-flooding, and do not add much system overhead. By exploring the content distribution, SAC drastically reduces the cost of similarity-based k-NN retrieval in P2P networks. The performance of SAC is evaluated through simulation study and compared against several search schemes as advanced in the literature.

1. Introduction

In recent years, peer-to-peer (P2P) networks are becoming popular in providing the ability of sharing data sources at a large scale. Conceptually, a P2P network is a collection of cooperative nodes that communicate with each other without the intervention of centralized indexing servers. These nodes are capable of not only storing and processing data, but also performing complex operations through their communications, such as P2P lookup [18] or multimedia data streaming [19]. Although many of the previous research focuses on the path finding protocols

that adapt to the dynamic network topology, information retrieval is becoming a hot issue in the applications of P2P networks. From the viewpoint of information retrieval, the organization of P2P networks can be classified into three categories:

Some earlier P2P networks, such as the Napster, are linked to centralized data source nodes (data centers) that host constantly updated directories of data contents. Queries issued from the client nodes are resolved at the data centers and the results are forwarded back to the requesting nodes through unicasting. Such centralized organization does not scale well and has the single points of failure. Moreover, the data center behaves as a hotspot and its data updates could increase the network traffic.

The more recently proposed P2P network frameworks are decentralized and have no data centers. The most commonly used frameworks are unstructured P2P networks, where the nodes form peer-to-peer connections among them and resolve queries through the cooperations with peers. Flooding is the most common approach for information retrieval in such P2P networks, since the requesting node does not have any information of the data contents of other nodes and has to employ the blind search. However, the flooding approach achieves good performance only when dealing with text information [21] due to its drastic consumption of system resources — storage, bandwidth, and energy. Considering the sheer size of the multimedia data, the performance deterioration is more drastic. In addition, the flooding strategy may cause duplicated queries and retrieval results, which may further increase the cost of the query processing.

To overcome the shortcomings of the blind search, the structured P2P networks were proposed in the recent literature as an alternative framework [20]. In such networks, the data objects are placed not at random nodes but at specified locations that will make

* This work is supported in part by National Science Foundation under contract IIS-0324835.

subsequent queries easier to satisfy. Moreover, the topology of such networks is strictly controlled and does not change drastically. Such designs improve the efficiency of information retrieval in some cases; nevertheless, at the cost of sacrificing the flexibility and scalability of the P2P networks. In practical applications, the network topology and the data contents of the nodes are constantly changing, which increase the difficulty of efficient data retrieval.

Due to the aforementioned reasons, P2P networks cannot utilize classical content-based retrieval methods that are based on centralized or flooding mechanisms. To overcome this difficulty, this paper describes a semantic-aware caching scheme (SAC) that facilitates k-NN retrieval in unstructured P2P networks. We address the fundamental problem of supporting nearest-neighbor retrieval within the network, in which each node caches a semantic content of earlier queries. By analyzing the cache content, an overview of data distribution in the network is obtained, and the later queries are resolved with optimized search cost.

The rest of this paper is organized into five sections: Section 2 introduces the background knowledge and related work. Section 3 outlines the preliminary concepts and the caching rationale. Section 4 evaluates the proposed scheme using experimental analysis. Section 5 draws the paper into conclusions.

2. Background and related work

2.1. Similarity-based retrieval

For years, similarity search on numeric data has attracted considerable research interest, especially in the multi-dimensional spaces, e.g. image feature space. The queries of semantically similar data are performed by conducting nearest-neighbor retrieval in the multi-dimensional spaces. The similarity search approaches that have advanced in the literature can be categorized into three classes: partition-based approaches, region-based approaches, and annotation-based approaches.

The partition-based approaches recursively divide the multidimensional spaces into disjoint partitions, with clustering or classifying algorithms, while generating a hierarchical indexing structure on these partitions. The earlier models in this class include quad-tree [8], k-d-tree [9], and vp-tree [9]. The recent research has focused on models based on clusters [7,10]. The partition-based approaches normally employ very complex computations, which make them inefficient for real-time data retrieval applications.

The region-based approaches employ small bounding regions (either in the form of minimum

bounding rectangles or spheres) to cover all the data objects. Based on these bounding regions, a balanced tree is constructed as the indexing structure. This class of indexing models includes the R-tree and its variations (R*-tree, R+-tree) [11], and SR-tree [12]. Relative to partition-based models, this indexing model improves storage utilization by avoiding forced splits. However, the data objects grouped in the same region may not share common semantic contents. Moreover, the performance of the region-based approaches degrades rapidly due to the existence of dimensionality curse [10].

The annotation-based approaches add some attached information to the data objects to facilitate similarity search. Gionis et al. [22] studied the relevance problem of similarity search for set data by embedding them into binary vectors in Hamming space using Min Hashing technique and error correcting codes. Sarwar et al. [23] used relevance feedback to improve the accuracy of similarity comparisons. Aggarwal et al. [24] propose the Signature table to map multi-dimensional data into strings called Super-coordinates, which are used as indexing entries. Tousidou et al. [25] and Nanopoulos et al. [26] proposed tree-structure indices called Signature trees, which employ bitmap signature to indicate the content relationships between parent/child nodes in the indexing trees. The annotation-based approaches achieve high accuracy in similarity search, and are efficient in handling data of different dimensionalities. However, these approaches are still based on centralized systems, and cannot be directly applied to wireless P2P networks.

In a P2P network with heterogeneous distributed data sources, traditional information retrieval methods, i.e., centralized or flooding strategies, may not work efficiently mainly because of the changes in network topology.

- Due to the heterogeneity and node mobility, the centralized search strategy is not an efficient choice in performing similarity search. Moreover, this strategy may also cause single point of failure, which results in low robustness.
- The flooding search strategy achieves good performance only when dealing with text information. Due to the sheer size of the multimedia data, the flooding strategy may drastically consume system resources.

2.2. Content distribution management

The goal of content distribution management is to provide an overlaid framework for the organization of P2P nodes. Many early proactive search protocols

depend on stationary routing tables to maintain route information between nodes. However, most practical P2P networks have dynamic topology and considerable overhead is generated in the maintenance of the frequently broken routes. To solve this problem, the reactive protocols employ on-demand discovering method to find the routes. However, the route may break as soon as they are discovered, wasting the bandwidth without getting any data across. Moreover, these methods fail to integrate the information of data contents into the process of finding routes, which increases the complexity of content-based retrieval.

A great deal of research has been done on organizing P2P networks based on content distribution. Search algorithms based on landmark hierarchies were proposed in [27]. The hierarchy consists of data nodes and landmark nodes that behave as indices. The landmark nodes self-organize themselves into a hierarchy, such that the landmarks at each level of the hierarchy show the approximate number of hops between the landmark nodes. Each data node maintains the shortest route to a landmark node in the hierarchy. Improvements to better organize the node contents and distance information by clustering the content-similar nodes and mapping them into semantic categories in the semantic space — such as Semantic-Aware Indexing Hierarchy [21]. However, topology / content changes may require costs to update such indices. Further improvements to search efficiency have led to continuing overlaid infrastructures (e.g. Chord [18]) that use content/location mapping to direct the searches to specific nodes holding the requested data. This approach causes higher maintenance overhead in the process of updating the mapping relationships in accordance with content changes.

Another type of solutions is based on the exploitation of P2P infrastructures. Content distribution paradigms such as Akamai and web caching hierarchies are often used to alleviate the work load on popular data servers. These networks are deployed by the content providers (i.e. servers), and web caches are usually deployed by clients. A content distribution network accepts the queries delivered from clients, and is connectionless and best-effort in nature. In such a network, nodes are not assigned to specific addresses, nor are queries addressed to specific nodes. Each node broadcasts a predicate indicating its contents and interest, i.e. the type of data and queries that it intends to receive. Similar frameworks also include the data sharing graph that dynamically adjusts the connectivity of each node based on its popularity.

3. P2P similarity search

3.1. Problem statement

We consider a P2P network where each node has a set of data objects to share with other nodes in the network. Each data object x is represented as a n -dimensional semantic vector $\varphi_x = (\omega_{1x}, \omega_{2x}, \dots, \omega_{nx})$. The attributes in the semantic vector could be manually added keywords, automatically extracted features, or relevance feedback from users. The P2P network can be considered as an undirected connected graph $G = (N, C)$, where $N = \{n_1, n_2, \dots, n_r\}$ is the set of nodes and $C = \{c_1, c_2, \dots, c_s\}$ is the set of wireless connections between nodes. Each node n_i may have a collection of data objects $x_1^i, x_2^i, \dots, x_r^i$, denoted as the node content $\chi(n_i)$. A data object x_j may exist in more than one nodes (i.e. replications), which are collectively denoted as a node set $\psi(x_j)$.

Without loss of generality, we assume the data objects $X = \{x_1, x_2, \dots, x_m\}$ are represented as data points in a n -dimensional semantic space R^n . The semantic similarity between two data objects is defined based on the Euclidean distance between their corresponding data points in R^n . Formally, the distance between two data objects x_i and x_j is defined as a cosine distance function $dist(x_i, x_j) = \cos^{-1}$

$$\frac{\varphi_{x_i} \bullet \varphi_{x_j}}{\|\varphi_{x_i}\|_2 \|\varphi_{x_j}\|_2},$$

where $\|\cdot\|_2$ denotes the Euclidean

vector norm.

- **Similarity query**

A similarity query (k -NN query) in a P2P network can be described as follows: Given the set of data objects $X \subset R^n$ and the P2P network G , for a specific integer constant k and a given query object x_q , return k data objects $X^* = \{x_1, x_2, \dots, x_k\} \subset X$ such that for $\forall (x \in X \wedge x \notin X^*)$ satisfies for $\forall i \in [1, k], dist(x_q, x) \geq dist(x_q, x_i)$.

In the context of P2P network, the resolution of k -NN query means flooding in the whole network and making pair-wise comparison on data objects in each node. The cost of this resolution is formidably high and impractical for real applications. Therefore, alternative approaches should be devised to perform the similarity search with optimized search operations and reduced cost.

Definition 1: Nearest-Neighbor Retrieval (k -NN)

Given an object set $X = \{x_1, x_2, \dots, x_m\}$ and a query object x_q , the nearest-neighbor retrieval of x_q within X , denoted as $\Xi^k(x_q, X)$, is the following set:

$$\mathcal{E}^k(x_q, X) = \{x_i \mid \forall x \notin \mathcal{E}^k(x_q, X), \text{dist}(x_q, x) \geq \text{dist}(x_q, x_i) \wedge |\mathcal{E}^k(x_q, X)| = k\} \quad (1)$$

Definition 2: Range-constrained k-NN search

Given a data object set $X = \{x_1, x_2, \dots, x_m\}$, a query object x_q , and a distance threshold d , the range-constrained k-NN search returns a set:

$$\mathcal{E}_d^k(X, x_q, d) = \{x_i \mid \forall x \notin \mathcal{E}_d^k(X, x_q, d), \text{dist}(x, x_q) \geq \text{dist}(x_i, x_q) \wedge \text{dist}(x_i, x_q) \leq d \wedge |\mathcal{E}_d^k(X, x_q, d)| = k\} \quad (2)$$

The significance of range-constrained k -NN search is to express the exact match k -NN queries into spatial range queries. By restricting the search range with the sphere centered at x_q with a radius d , the search cost can be drastically reduced. In addition, it can be proven that the range-constrained k -NN search returns exactly the same query result as the original k -NN search. The reason is that the data objects with closest semantic distance with x_q are located within a sphere centered at x_q , by selecting appropriate distance threshold, these data objects can be found through a range-constrained k -NN search operation.

Claim 1: Given a multimedia data object set $X = \{x_1, x_2, \dots, x_m\}$ and a query object x_q , there always exists a distance threshold d , satisfying $\mathcal{E}^k(x_q, X) = \mathcal{E}_d^k(x_q, X, d)$.

Proof: For a given distance threshold d , the multimedia data object set X is divided into two partitions — the objects whose semantic distance to x_q is less than d (*i.e.* within the sphere centered at x_q with a radius d) and the objects outside the sphere. By adjusting the distance threshold, a sphere that encloses exactly the same data objects in $\mathcal{E}^k(x_q, X)$ can be obtained. In other words, given the multimedia data object set X , the distance threshold can be considered as a function whose parameter is the number of nearest neighbors. Therefore, the threshold d satisfying $\mathcal{E}^k(x_q, X) = \mathcal{E}_d^k(x_q, X, d)$ always exists. ■

• **Search cost**

A similarity-based k -NN query may travel several nodes before reaching the data source nodes containing the requested query result. A query may also be decomposed into multiple sub queries and forwarded to different nodes for query resolution. The data objects collected from different network branches are then merged together as the result of the original query. Therefore, we evaluate the query resolution in a P2P network from three aspects: search cost, accuracy, and system overhead. The search cost includes query response time, message complexity per query, and cache hit ratio when caches are employed. The

accuracy means the percentage of the results generated by decentralized search strategy matching with the results generated by centralized search strategy. The system overhead includes the maintenance cost (messages and time) incurred in the process of adjusting the search index or cache content in accordance with the network topology and data content changes.

A desirable system should guarantee high search effectiveness with low search cost and system overhead. It should also provide scalability and robustness to network size and query frequency. Due to the space constraint, we focus on the aforementioned three aspects as the major performance metrics of similarity search.

3.2. Proposed scheme

In this section, we introduce a multi-scale regression graph (MSRG) to represent the semantic relationships among multimedia data objects. The goal of our model is to provide a paradigm that avoids the semantic gap by employing the linguistic-aware semantics in the data content representation. The literature has reported works on 2D hidden Markov model (HMM) [28], which consider the feature vectors statistically dependent through an underlying Markov mesh, indicating the conditionally transitions between image blocks in both horizontal and vertical directions. As an approach extending the representation to multimedia data objects, our MSRG model tries to use the states to represent the semantic categories, and describes the relationships between semantic categories using probabilistic transitions.

• **Semantic regression graph**

The semantic regression graph is a multi-level HMM in which each level indicates a semantic granularity in the semantic space. Suppose there are N states $\{1, \dots, N\}$ and the transition probability between state i and j is defined as $\lambda_{i,j}$. Unlike the Markov mesh, our MSRG only allows the transitions between states in neighboring levels. The idea is to represent the semantic categories in the application domain using the states, denoted as $\epsilon_1, \epsilon_2, \dots, \epsilon_N$. Suppose the semantic categories can be divided into M levels (based on heuristic information), for the simplicity of presentation, the states in the MSRG are denoted as $\{\{\epsilon^1_1, \epsilon^1_2, \dots, \epsilon^1_S\}, \dots, \{\epsilon^M_1, \epsilon^M_2, \dots, \epsilon^M_T\}\}$.

The MSRG model is a uni-directional weighted connected graph. The weight on each edge indicates the probability of transition between the two states. For any given state ϵ^n_i on level n of MSRG, there are two

types of edges (upward edges and downward edges) connecting it with the states in neighboring semantic levels. Figure 1 illustrates an example of the transition edges between states in two levels.

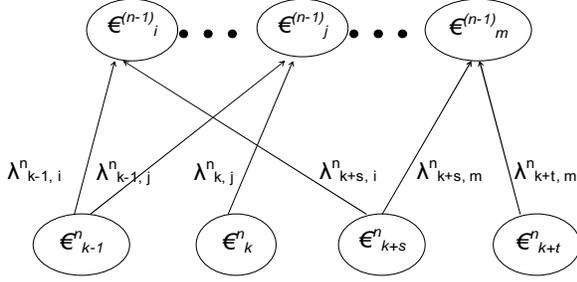


Figure 1: The semantic relationships between states.

- **Semantic space partitioning**

Given a set of semantic categories $\{\{\epsilon^I_1, \epsilon^I_2, \dots, \epsilon^I_S\}, \dots, \{\epsilon^M_1, \epsilon^M_2, \dots, \epsilon^M_T\}\}$, the lowest level $\{\epsilon^M_1, \epsilon^M_2, \dots, \epsilon^M_T\}$ represents the basic semantic categories that can be further summarized by other categories. Suppose there exists a semantic level satisfying that each node in this level (say ϵ^K_i) completely covers a collection of basic semantic categories $\{\epsilon^M_1, \epsilon^M_2, \dots, \epsilon^M_T\}$ (i.e. all upward paths starting from these basic categories must go through an individual node in ϵ^K_i). In such a case, the semantic level K is considered as a “cut” of MSRSG, and the semantic categories in level K form a partition of the semantic space which includes orthogonal sub regions of the space. The diagonal length of the sub region of ϵ^K_i is called the size of the semantic category, denoted as $\Phi(\epsilon^K_i)$. In the range-constrained k-NN search

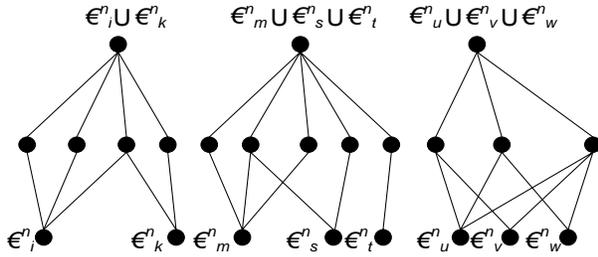


Figure 2: The enclosure of semantic categories.

we use the smallest size of the semantic categories in a cut to determine the distance range of k-NN. The semantic content of ϵ^K_i is determined by the basic semantic categories in its enclosure, equivalent to the union of the basic categories.

- **Semantic-aware caching**

The aforementioned definitions depict the data objects as a set of points in the semantic space, whose contents can be collectively described using regression graph of semantic categories. Basing on this representation method, we propose to cache the constraints as the concise description of query results, with the aim of increasing cache hit ratio and reduce query resolution cost.

Logically, the local cache of a node n_i is divided into a set of cache entries — each entry indicates one or multiple nodes in the network. A cache entry is a triple (hit region, miss region, node list). The hit region is the constraint-based description of resolved queries, which can be considered as the semantic categories covering the data points of earlier query results. The miss region shows the unresolved queries, which can be represented as the sphere sub space where no query results are found. The node list shows the nodes whose data contents can be characterized by the hit region and the miss region. Figure 3 illustrates the cache structure.

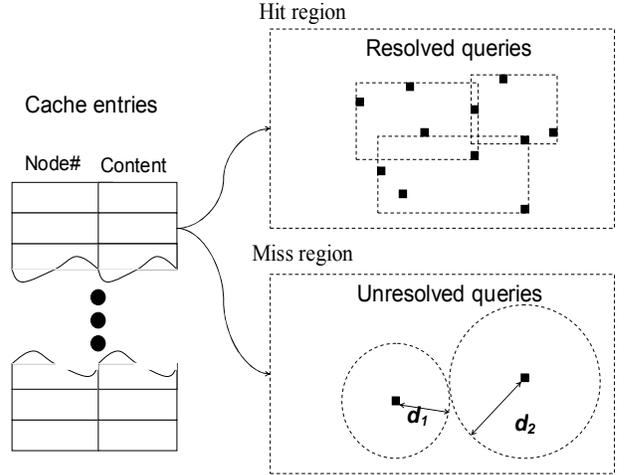


Figure 3: An illustrative example of cache entries.

Physically, we store the content descriptions using a hierarchical organization of data contents. The semantic categories and distribution density, in the form of state ids and logic constraints, are stored in one or multiple linked cache entries. Note that there are three possible relationships between the sub spaces described by the constraints: enclosure, overlapping, and isolation. Basing on these relationships, a hierarchical indexing structure can be built on the cache entries, which maintains the semantic descriptions as well as the physical storage information for every cache entry.

4. Performance study

The evaluation consists of a series of experiments conducted using both real world data sets and simulated environments. The purpose of experimental analysis using both real data and synthetic data is to evaluate the performance gain of the proposed caching scheme in the context of different network scales and data sources. Our comparative analysis is based on various performance metrics such as accuracy, search cost, scalability, and physical characteristics of the mobile nodes.

The real world data set comprises up to 4630 images of 110 semantic categories from the Corel dataset. To examine the effect of semantic locality, we randomly select 400 images as a “hot” dataset, and let the probability of queries on the hot dataset abiding a given parameter.

We compared and contrasted the performance of our MSRSG-based caching with two recently proposed P2P caching models — hierarchical spatial caching (HSPC) [29] and proactive caching (PROC) [21].

The accuracy of the caching models is evaluated as the number of semantic categories needed in the cache contents to achieve certain accuracy. Given the real world image dataset including the 110 basic semantic categories, Table 1 lists the needed categories for some accuracy settings. As one can conclude, MSRSG needs the smallest number of categories, and therefore needs less cache storage space.

Table 1: The comparison on accuracy.

Accuracy	20%	25%	30%	40%	50%
MSRSG	2	4	7	9	12
HSPC	9	17	31	46	63
PROC	13	29	47	65	97

In a P2P network, the cache hit ratio is often used to evaluate the performance of caching approaches. Traditional caching schemes rely on large caches and complex replacement policies to achieve high hit ratio, while MSRSG-based caching uses the relationships on the semantic categories to reduce space requirement and improve hit ratio. In another simulation run, we used the synthetic data set of 10,000 data objects disseminated on 1,000 P2P nodes. The result is shown in Figure 4. As one can see from the figure, all three caching models improve hit ratio as less queries are issued in a time unit, and MSRSG achieves the highest hit ratio due to its less requirement of cache space.

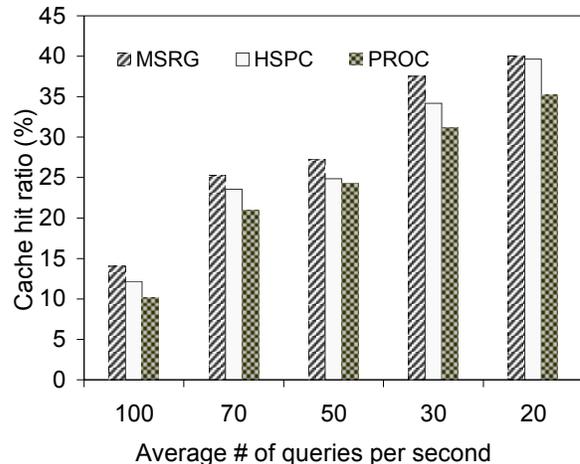


Figure 4: The comparisons on cache hit ratio.

5. Conclusions

We proposed a dynamic semantic-aware caching scheme that facilitates content-based multimedia retrieval in peer-to-peer networks. This scheme is based on analysis of cached query results to represent the data contents in each node. It has several innovative characteristics such as content distribution representation and non-flooding query processing.

The proposed scheme makes use of the data content distribution in P2P networks to reduce the search cost of k -NN queries without incurring high maintenance overhead. Through extensive performance analysis, we found that the semantic-aware caching methodology has the following features:

- The SAC is a decentralized non-flooding strategy facilitating content-based multimedia retrieval in P2P networks. As shown in our simulation results, it can achieve high hit ratio while visiting only a small portion of nodes.
- We employed content distribution information in the organization of cached data — the k -NN queries are forwarded only to the content-related nodes.
- Our model is dynamic and capable of self-organizing itself as the network status changes. This further offers scalability and robustness in large-scale networks.

6. References

- [1] C. E. Perkins, E. M. Royer, and S. R. Das. Performance comparison of two on-demand routing protocols. *IEEE Personal Communications*, 2001.

- [2] J. Broch, D. Maltz, D. Johnson, Y. Hu, and J. Jetcheva. A performance comparison of multi-hop wireless ad hoc network routing protocols. *ACM Mobicom*, 1998: 85-97.
- [3] L. Yin and G. Cao. Supporting cooperative caching. *IEEE INFOCOM*, 2004.
- [4] A. Gionis, D. Gunopulos, and N. Koudas. Efficient and tunable similar set retrieval. *Proc. ACM Sigmod* 2001.
- [5] B. Sarwar, G. Karypis, and J. Riedl. Analysis of recommendation algorithms for e-commerce. *ACM Conference on Electronic Commerce 2000*, pp 158–167.
- [6] J. He, M. Li, H. Zhang, H. Tong, C. Zhang. Manifold-ranking based image retrieval, *ACM Multimedia*, 2004.
- [7] M. Swain and D. Ballard. Color indexing. *Int. Journal of Computer Vision*, 7(1):11-32, 1991.
- [8] H. Samet. The Quadtree and related hierarchical data structures. *ACM computing surveys*, 1984: 187-260.
- [9] A. W. Fu, P. M. Chan, Y. Cheung, and Y. S. Moon. Dynamic VP-tree indexing for n-nearest neighbor search given pair-wise distances. *VLDB 2000*: 154–173.
- [10] K. Beyer, J. Goldstein, and U. Shaft. When is "nearest neighbor" meaningful? *VLDB*, 1994:487-499.
- [11] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles, *Sigmod*, 1990:322-331.
- [12] N. Katayama, and S. Satoh. The SR-tree: An Index Structure for High-dimensional Nearest Neighbor Queries, *Sigmod*, 1997, 26(2):369-380.
- [13] X. Tang and J. Xu. On replica placement for QoS-aware content distribution. *Infocom*, 2004.
- [14] E. Tousidou, A. Nanopoulos, and Y. Manolopoulos. Improved methods for signature tree construction. *Journal of Computing*, 2000.
- [15] Q. Ren and M.H. Dunham. Using semantic caching to manage location dependent data in mobile computing. *ACM Mobicom*, 2000: 211-221.
- [16] T. Hara. Efficient replica allocation for improving data accessibility. *Infocom*, 2001.
- [17] C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-peer information retrieval using self-organizing semantic overlay networks. *ACM Sigcomm*, 2003:175-186.
- [18] I. Stoica, R. Morris, D. Karger, M. Kaashock, and H. Balakrishman. Chord: A scalable peer-to-peer lookup protocol for internet applications. *ACM SIGCOMM*, 2001.
- [19] G. Auffret, J. Foote, C. Li, B. Shahraray, T. Syeda-Mahmood, and H. Zhang. Multimedia access and retrieval: The state of the art and future directions, *ACM Conference on Multimedia*, 1999: 443-445.
- [20] V. Dheap, M. Munawar, and S. Ward, Parameterized neighborhood based flooding for ad hoc wireless networks. *IEEE Milcom*, 2003: 1048-1053.
- [21] B. Yang and A. R. Hurson. Supporting Semantic-Based Multimedia Data Access in Ad Hoc Networks. *IEEE Wowmom*, 2005.
- [22] A. Gionis, D. Gunopulos, and N. Koudas. Efficient and tunable similar set retrieval. *Proc. ACM Sigmod* 2001.
- [23] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. *Proc. ACM Conference on Electronic Commerce 2000 (EC'00)*, pp 158–167.
- [24] C. Aggarwal, J. Wolf, and P.S. Yu. A new method for similarity indexing of market basket data. *Proc. ACM Sigmod*, 1999.
- [25] E. Tousidou, A. Nanopoulos, and Y. Manolopoulos. Improved methods for signature tree construction. *Journal of Computing*, 2000.
- [26] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. A data mining algorithm for generalized web prefetching. *IEEE Transaction of Knowledge and Data Engineering*, 2002.
- [27] G. Pei, M. Gerla, and X. Hong, LANMAR: landmark routing for large scale wireless ad hoc networks with group mobility, In *Proceedings of the 1st ACM international symposium on mobile ad hoc networking & computing*, 2000, 11-18.
- [28] J. Li, A. Najmi, and R. M. Gray, Image classification by a two dimensional hidden Markov model, *IEEE Transaction on Signal Processing*, 2000, 48(2), pp 517-533.
- [29] R. Schmidt, B. Wyvill, E. Galin, Interactive implicit modeling with hierarchical spatial caching. *Proceedings of Shape Modeling International*, 2005, pp. 104 – 113.