Prefetching: A Survey

[1]Dr. M. Thangaraj, [2]V. T. Meenatchi
[1]Madurai Kamaraj University, Madurai, Tamil Nadu, India,
[2]Thiagarajar College, Madurai, TamilNadu, India.

Abstract
Current Web is a huge repository, with its explosion in both the range and quantity of web content. Even, when the speed of Internet has improved with the reduced costs, the traffic is getting heavier. The enormous information makes it difficult to find the relevant information quickly. This led to the effort to improve the speed, by reducing the latency, make the web more relevant and meaningfully connected. Prefetching is the method for reducing Latencies. This area of research gains importance since, an user always expects an interactive response, better satisfaction and quality of output. This paper presents the various issues in Prefetching.

Keywords
Semantic Web, Prefetching, Cache, Latency.

I. Introduction

The principle technique used to reduce memory access time is the use of some form of Cache hierarchy. Caching is the storage of recently retrieved computer information for future reference. A request for an object found in the cache is called cache hit, otherwise cache miss.
Prefetching is used as an attempt to place data close to the processor before it is required, eliminating as many cache misses as possible. Caching offers the following benefits: Latency reduction, Less Bandwidth consumption, Lessens Web Server load. Prefetching is the means to anticipate probable future requests and to fetch the most probable documents, before they are actually requested. It is the speculative retrieval of a resource into a cache in the anticipation that it can be served from the cache in the near future, thereby decreases the load time of the object.
Semantic Web is the next generation of the Web, which provides Semantic annotations to the web contents and describes the meaning of the content. Various research issues on Prefetching have been identified. This paper mainly concentrates Prefetching into three major directions: Web Prefetching, Data Prefetching and Other issues (fig. 1), Web Prefetching is further classified into Cache Prefetching, Proxy Prefetching and Semantic Prefetching, Data Prefetching is classified into Content Prefetching and Context Prefetching.

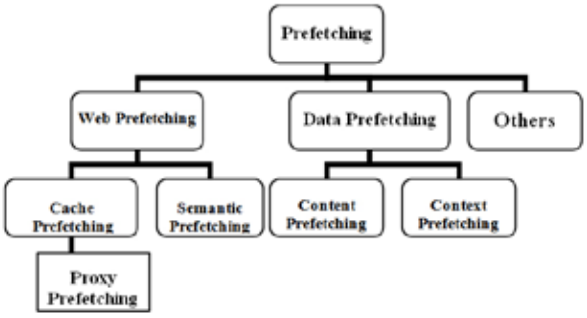

Fig. 1: Prefetching Tree

The rest of the paper is organized as follows, Section 2 discusses the Web Prefetching, Section 3 describes Data Prefetching and Section 4 deals with Other directions related to Prefetching, Section 5 provides some Conclusion remarks.

II. Web Prefetching Techniques:

The basics of web prefetching techniques preprocesses the user requests, before they are actually demanded. Therefore, the time that the user must wait for the requested documents can be reduced by hiding the request latencies [27]. Various approaches and algorithms have been proposed for improving the web performance.

The Markov prefetcher [10] reduces the overall execution stalls due to data and memory operations. The prefetcher might be useful, when it observes the reference sequence prior to the predicting references, that have not occurred. The drawback of the prefetcher is, it does not launch multiple prefetch requests and prioritize them and it must consider some mechanism to limit the bandwidth devoted to prefetching.

As an effort to improve predictive pre-fetching precision and reduce user's access time, Hidden Markov Model prefetching model [29] was identified. It captures and mines the latent information requirement concepts that the user's access path contains and to make Semantic based prefetching decisions. This approach as such cannot be applied in Information Filtering, Information recommender system and personal website.

The web object prefetcher [2] implements a web-application centric Markov model with a Knapsack object selector that adapts to a client's web-application access pattern. The Knapsack object selector supplements the prefetch technique in identifying objects and provides the greatest profit, consequently increasing the object cache-hit percentage. However improving the model's performance and its prediction accuracy is the area of ongoing research.

An integration of Semantic information from an underlying ontology into probabilistic low order Markov model [15] is proposed, to overcome the problem of contradicting prediction. This work does not gather more semantic information. This when gathered, aids in inferencing along with association rules in the recommendation/ prediction phase of web usage mining.

A few researchers have used Prediction by Partial Match (PPM) based models for web prediction. Anticipating the user's next requests by extracting useful knowledge from historical user requests is done through Prediction by partial Match (PPM). Online PPM model is implemented based on non-compact suffix tree [33]. The prediction model combines entropy, prediction accuracy and the longest matching rule.
This method gives a good accuracy only when there is a longest repeated sequence to predict the next request. A novel PPM model based on stochastic gradient descent [34] combines all the features like page access frequency, prediction feedback, context length and conditional probability, which greatly influences the prefetching performance of PPM models and reduces the incorrect predictions.
A Stochastic Petri Nets (SPN) based integrated web prefetching and caching model[14] considers hit ratio and byte hit ratio in addition to access latency, throughput. Further research work [13]

includes knowledge prefetching based on the knowledge grid, which provide the basis for more intelligent prefetching. The fig. 2 shows a SPN model for Web prefetching and caching in which p1 stands for client, p2 for cache, p3 for normal requests, p4 for prefetching requests, p5 is for server, to process the normal requests, p6 is for the server, to process prefetching requests, p7 for the idle state of server.

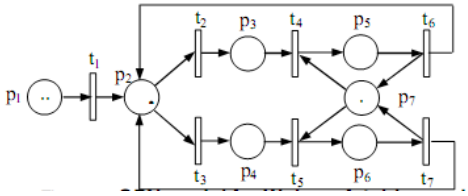

Fig. 2: SPN Model for Web Prefetching and Caching.

The firing of t1 stands for the arrival of user requests to local cache, t2 for the users, to issue normal requests for objects on the shared server, t3 for the users, to issue prefetching requests for objects on the shared server, t4 for the normal requests to the server, t5 for the prefetching requests to the server, After the server's processing, objects are download to the local cache at t6, After the server's processing, prefetched objects are download to local cache at t7.

From the point of view of SPN theory, the access latency of passive caching is considerably reduced. However the drawback is that, if p increases, the access latency increases. The reason is that the number of prefetching objects decreases, which can avoid consuming too much of network resources.

During the view time of the current web page, its associated hyperlink set is used to predict which object to prefetch. A Semantic Link Prefetcher[3] is meant for this purpose. The use of semantic links that connects documents has been shown to improve the performance of search engine accuracy, by reducing the rendering delays associated with web-applications and slow communication channels. The following Table describes the various Semantic Link types and its definition.

Table 1: Semantic Link Types

Link Type	Definition
Sequential link (seq)	The predecessor web page should be browsed or used before the successor web page.
Similar-to link (simi)	The semantics of the successor web page are similar to those of its predecessor web page.
Cause-effective link (ce)	The predecessor web page is the cause of its successor web page; the successor is the effect of its predecessor.
Implication link (imp)	The semantics of the predecessor web page imply the successor web page.
Subtype link (st)	The successor web page is a part of its predecessor web page.
Instance link (ins)	The successor web page is an instance of the predecessor web page.
Reference link (ref)	The successor web page is a further explanation of the predecessor web page.

The results exhibit a significant increase in the cache-hit percentage, which indicates that objects are being acquired before they are requested. However the cost associated with augmenting hyperlinks with semantic information is minimal, this will be further simplified as web design tools that incorporate the ability to indicate a semantic label to a hyperlink as an additional attribute like font color, style etc.

A new prefetching algorithm based on the Adaptive Resonance Theory (ART) [26] of neural networks is proposed in this work. The new model uses the bottom-up and top-down weights of the cluster-URL connections obtained from a modified ART1 algorithm to make prefetching decisions. Each URL and clusters are connected together with bottom-up and top-down weights respectively. In Fig. 3, each URL is connected to the m clusters.

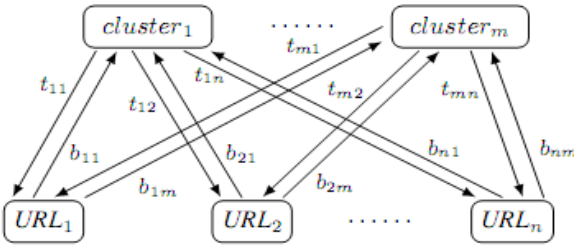

Fig. 3: Top Down and Bottom Up Weights

The advantage of using the ART algorithm is that it adapts to the change in users Web access patterns over time without losing information about their previous web access history. With this approach, each cluster of users is represented by a prototype vector that is a generalized representation of URLs frequently accessed by all the members of that cluster. The algorithm aims particularly for the digital library environment. The clustering based approach is selected to match the user and the document categories. Further the algorithm can be extended to link the document's abstract and references by adjusting the extraction of pattern vectors. This work does not support non-cluster based applications.

This paper reviews the popular prefetching algorithms [4] based on Popularity, Good Fetch and Lifetime. A family of algorithms intended to improve the respective prefetching performance metrics under consideration - the hit rate, the bandwidth, or the hit ratio. Within the family of Objective-Greedy algorithms, the Hit Rate-Greedy and Bandwidth-Greedy prefetching achieve the highest hit rate and lowest bandwidth. A strict optimal algorithm must be obtained by a solution to the maximum weighted average problem with pre-selected items.

As a continuation of the previous work, the Hit rate and Bandwidth(H/B) Optimal prefetching algorithm that obtains the maximum H/B ratio is proposed. This algorithm [5] provide better objective based performance. Given the total number of objects, it may be a challenge to find the n-max value for H/B Greedy and H/B Optimal and the future work is to extend this algorithm in wireless networks and P2P networks.

A. Cache Prefetching:

Web caching and prefetching [18,28] have been recognized as the effective schemes to alleviate the service bottleneck, minimize the user access latency and reduce the network traffic. These two techniques are very useful tools to reduce congestion, delays and latency problems.

A new ODBMS prefetching framework has been proposed, Path and Cache Concious Prefetching. This framework [32] creates new prefetching algorithms and it outperforms the existing prefetching algorithms like PPM-1, PPM-3 and PCM in terms of less stall time for small buffer sizes, uses less storage overheads. The future work may proceed with how clustering statistics can be used by prefetching algorithms to improve the performance.

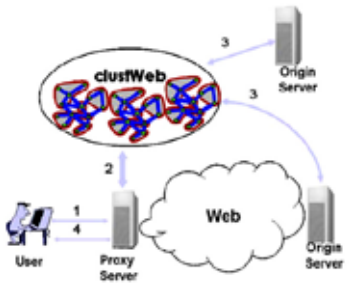

Fig. 4: The ClustPref scheme

As a continuation of the previous work, a clustering-based prefetching scheme [11], which uses a graph-based clustering algorithm was devised. It identifies the clusters of ''correlated'' web pages based on the user's access patterns. This scheme can be integrated easily into a web proxy server, improving its performance. This clustPref clustering scheme based on web prefetching is shown in fig. 4.

A web user requests an object in a particular time. The proxy identifies the web user, according to its IP address, and assigns it to one of the client groups. The proxy searches the requested web object inside the existing cluster. The proxy prefetches all objects exist in the cluster, stored at the origin server. Then, these objects are managed by the proxy's cache replacement policy and the proxy sends the requested object to user.

This scheme have practical applications in Information Management and e-commerce only. It does not focus on the wide range of applications, such as discovering usage patterns and profiles, detecting copyright violations, and reporting search results and Content Distribution Network.

A prefetch cache sizing module for use with any sequential prefetching scheme was proposed in [25]. Disk array caches perform sequential prefetching by loading data contiguous to I/O request data into the array cache. The simple sizing module can be added to any prefetching scheme to ensure that the prefetch cache size is adequately matched to the requirement of the prefetching scheme on a dynamic workload comprising multiple streams. Even when, it achieves the maximum hit rate, the drawback of the sequential prefetching technique, is that it generates only the prefetch requests for data that are stored contiguously to the I/O request data.

Proxy Server Prefetching: Proxy Server plays a major role in prefetching. An Intelligent prefetching [20] reduces document retrieval latency at the expense of minimal increase in network traffic at the user end. A monitor agent predicts possible future requests that are then prefetched to a proxy server "near" the user. Objects are pre-fetched to the personal cache at the proxy server to get faster downloading. Since the personal caching agent is used, the probability of finding the objects at the proxy server is greater and fewer objects are downloaded from the Internet becomes an issue.

This paper [31] proposes a new approach for predicting user requests based on the recent behavior of the individual users from the access log of a proxy server. This approach derives patterns, organized as a compact index and aids in effective prediction. Even when the experiments prove the best hit ratio of the prediction achieves 75.69%, while the longest time to make a prediction only requires 1.9 ms. It did not focus on the prediction of web page accesses by proxy server log to higher service rate.

The client cache [2] is divided into a "regular" cache and "prefetching cache". The regular cache is meant for on demand requests and prefetching cache is for prefetched requests. A set of clients are connected to a proxy server through bandwidth limited dedicated lines (e.g., dialup phone lines). The proxy server implements its own caching system. The access time improvement is considered as the performance metric. The effect of the caching system and user "think time" on the effectiveness of prefetching was tested. A prefetching protocol based on this model does not focus on all the parameters of prefetching is a major drawback.

B. Semantic Prefetching:

Existing prefetching algorithms are mostly based on URL graphs, which is effective in prefetching of documents that are often accessed, few of them can prefetch documents whose URLs have never been accessed. The following three papers[6, 8-9] have been proposed to overcome the limitation of URL graph based approach. A semantics-based prefetching technique is applied in News Surfing [6]. It captures the semantics of a document by identifying keywords on its URL anchor texts. It relies on neural networks over the keyword set to construct the semantic-net of the URLs and predicts the user's future requests. This approach takes advantage of "semantics locality" of access patterns and is able to prefetch documents whose URLs are never been accessed before. It also features a self-learning capability and good adaptive to the change of user surfing interests. In addition to the news services, the prefetching technique could also be applied to other application domains, such as on-line shopping and auction, where semantics of web pages are often explicitly presented in their URL anchor texts. Such domain-specific prefetching facilities can be turned on or off on-demand. The cache would contain some prefetched documents that are never accessed by the client. This information will be used as a feedback to improve the prediction accuracy in the future. The architecture in fig. 5 assumes a personalized information agent (or proxy) is running behind the browser interface and watching over its client web access behaviors. In step 1, this agent analyzes the documents that its client is browsing with an intention to extract their semantics and find out the semantic relationships between the accesses. In the next step, the semantic information is represented in an open model, which accumulates client knowledge about the client preferences with a self-learning capability. Based on the knowledge, In step 3, the agent makes predictions about its client's future requests and "steal" network bandwidth to prefetch related Web documents from the Internet in step 4.

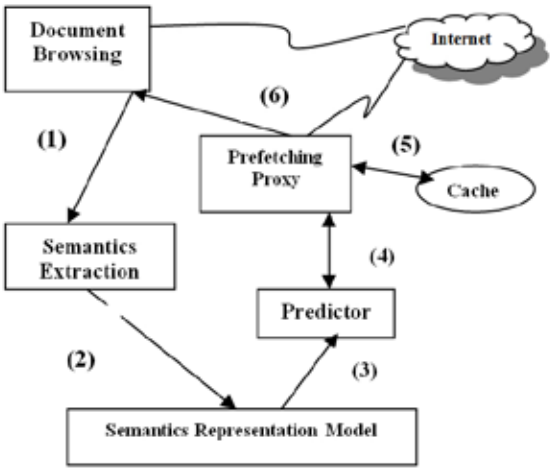

Fig. 5: Architecture of Semantics based Predictive Prefetching.

In the next step, the prefetched documents are temporarily stored in the agent's internal cache. This prefetching agent also serves as a client-side proxy of the browser. In the last step, it provides the browser with requested web pages transparently from its cache, if they are available. Otherwise, it fetches the pages from the Internet. The problem of automatic semantics extraction from the general HTML document is a major issue of research, to be resolved.

This Semantic based prefetching [8] technique predicts the future requests based on the semantic preferences of past retrieved document. This technique is applied in News Surfing. The system extracts document semantics by identifying keywords in their URL anchor texts and relies on neural networks over the keyword set to predict the future requests. The technique was implemented in a NewsAgent system and cross-examined in daily browsing of ABC.com, CNN.com, MSNBC.com news sites.

The results have showed 60% hit ratio due to the prefetching approach. Constructing a prediction unit using neuron networks, instead of single neurons is a major topic of research. This predictor considers only the keyword as a major factor, other factors such as document size and the URL graphs of a hypertext document, must be considered for the design of the predictor and the Cache utilization has to be still improved. Deployment of other statistical classification and learning techniques like Bayesian networks will be the future work, to be carried out.

A Keyword-based semantic prefetching [9] predicts future requests based on semantic preferences of past retrieved web documents. A client-side personalized prefetching system, NewsAgent was implemented in Internet News services. The system exploits semantic preferences by analyzing keywords in URL anchor text of previously accessed documents in different news categories. It employs a neural network model over the keyword set to predict future requests.

The NewsAgent system is based on keyword matching. Two issues related to keywords are synonymy and polysemy. This technique alleviates the impact of polysemy by taking advantage of the categorical information in news services. Most notable technique with respect to context information retrieval that addresses the two issues is, Latent Semantic Indexing, which infers the structure of relationships between article and words. However, with the popularity of XML or RDF compliant web resources, the semantics-based prefetching techniques can still be improved in reducing the web access latency, thereby providing more personalized quality services on the semantic web.

A Semantic Prefetching scheme [21] uses a new dissemination method for semantic information, using a dynamic caching strategy, that considers a reference of user in Semantic Prefetching Area (SPA). By applying semantic prefetching at SPA, the network traffic, response time, and communication cost are reduced. The server keeps log files represented by the semantic description and searches the prefetching description into log files.

A new query processing algorithm with a Preference Priority Replacement strategy to manage cache was introduced. This scheme has shown that this scheme is superior to existing caching mechanisms for location-based services in the aspect of data access efficiency and cache availability. The future efforts must emphasize on retaining the overall performance, which is greatly influenced by efficient query processing, considering dynamic user patterns.

Prefetching and caching [22] are exemplary techniques for data transmission, and offer advantages in user-centric services in location-dependent data environments. However, in mobile environments, prefetching and caching inevitably require frequent uplink requests because the data that is needed in the clients' current location has to be transmitted from a server.

To overcome this drawback, a semantic prefetching scheme, with descriptions and cache replacement policies, based on range query processing algorithms are presented. This scheme based on semantic prefetching can manage various user types with different mobility and query patterns.

III. Data Prefetching:

Data prefetching is a data access latency hiding technique, which decouples and overlaps data transfers and computation. In order to reduce CPU stalling on a cache miss, data prefetching predicts future data accesses, initiates a data fetch, and brings the data closer to the computing processor before it is requested. A data prefetching strategy has to consider various issues in order to mask data access latency efficiently. It should be able to predict future accesses accurately and to move the predicted data from its source to destination in time.

Various criteria influence the performance of prefetching strategies. This paper [24], provides the taxonomy of the five primary issues (what, when, destination, source, and initiator), that are necessary in designing prefetching strategies. A prefetching strategy is effective for multi-core processing environments, when it is adaptive to choose among multiple methods to predict future data accesses.

When a data access pattern is easy to found, prefetching strategy chooses history-based prediction algorithms to predict future data accesses. If data accesses are random, using the pre-execution based approach would be beneficial.

A. Content Prefeching

This dissertation [19] has presented content sensitive data prefetching, a new hardware only technique for data prefetching This study has shown that a simple, effective, and realizable content sensitive data prefetcher can be built as an on-chip component of the memory system, and will scale well with future processor designs. One of the major findings is that the content prefetcher enhanced memory system is capable of delivering timely prefetches to fully suppress demand load misses for nearly all applications.

The large percentage of full content prefetch hits fully validates the decision to place the content prefetcher on chip and positively indicates that the proposed memory system that utilizes prioritized memory traffic and prefetch reinforcement is well balanced.

B. Context based Prefetching:

A technique [17] for predicting useful prefetches when a navigational object-oriented interface was implemented on a relational DBMS. When implementing persistent objects on a relational database, a major performance issue is prefetching data to minimize the number of round-trips to the database. This paper proposes the use of the context in which an object is loaded as a predictor of future accesses, where a context can be a stored collection of relationships, a query result, or a complex object.

When an object O's state is loaded, similar state for other objects in O's context is prefetched. The future work in this area may provide the environment with object-relational DBMSs, and offer persistent object interfaces.

Since the memory access performance lack, far behind computational performance, data access delay has a severe impact on overall system performance [30]. The recent advances in multicore and many core processor architectures has insisted the steps for reducing data access delay for high-end computing. Data prefetching mechanisms, especially the general context-based prefetching approach, have received intensive attention and have been proven effective in masking the processor-memory performance gap.

An analytical study on comparing different orders of context-based prefetching models was done and a new multi-order context-based (MOC) data prefetching that incorporates multi-order context analysis to achieve better overall prefetching effectiveness was invented. Even when the MOC prefetching, achieves considerable latency reduction and is promising in hiding data access latency for high-end computing systems.

The selection of multiple orders, more intelligently and dynamically combine the multi-order context analysis and global context analysis, the correlation from all instructions to further improve context-based data prefetching is the future issue.

IV. Other Directions:

Prefetching is applied in various domains to improve the system performance. A location-aware prefetching [12] mechanism is introduced that is independent of any additional infrastructure and that gathers information solely over deployed, low-bandwidth wireless links. Location-awareness becomes more and more important for delivering relevant data. The latter is a challenging task when dealing with unstable and low-bandwidth wireless networks, especially in areas that are not or only poorly covered.

Prefetching is an elegant technique to handle information provisioning for users under these circumstances. In this paper, new strategies for prefetching in a location-aware surrounding are investigated. The mechanism, as well as the underlying location model, could however also be transferred to other application areas such as railway networks and highways. The dynamic service priority allocation and the development of location-aware cache invalidation algorithms are the focus of research in this area.

To overcome the problem of long retrieval latency [7] caused by the unpredictable user behaviors during multimedia presentation, a prefetching scheme using the association rules from the data mining technique was proposed. The data mining technique can provide some priority information such as the support, confidence, and association rules which can be utilized for prefetching continuous media. The prefetching policy can predict user behaviors and evaluate segments that may be accessed in near future. The results show that the prefetching approach can get better latency reduction, even for small cache size.

Mobile users connected to wireless networks expect performance [16] comparable to those on wired networks for interactive multimedia applications. Satisfying Quality of Service requirements for such applications in the wireless networks is a challenging problem, due to limitations of low bandwidth, high error rate and frequent disconnections of wireless channels. In addition, wireless networks suffer from varying bandwidth. This paper has presented an access prediction algorithm that uses a compound access graph with access-after semantics. Access improvement was used as a performance measure. This is expressed in terms of resources (retrieval time, viewing time and cache) and access probabilities.

The proposed access model can be used effectively to enhance a mobile user's access to wireless multimedia. The effect of compound prefetch on network load is the further issue to be resolved.

V. Conclusion

Lot of Research is going on in Prefetching, under various directions. In this work, three major areas: Web Prefetching, Data Prefetching and Other Directions are analyzed and discussed. Web Caching was the effective technique used to reduce the network latency, prior to Web Prefetching. Although web cache schemes, reduce the network load and bandwidth consumption, they still suffer from a low hit rate, stale data and inefficient resource management.

Web Prefetch schemes overcome the limitation of web cache mechanisms through pre-processing contents before a user request comes. The web prefetch schemes focus on the spatial locality of objects when current requests are related with previous requests. These schemes increase the bandwidth utilization and reduce or hide the latency due to bottleneck at web server.

Data Prefetching deals with access latency hiding. Prefetching techniques are applied not only to reduce the access latency, it is also applied to reduce the network traffic and improves the user satisfaction. It is applied in wireless networks and efficiently improves the data management. At present, the Semantic Prefetching gains its importance in various domains and opens a vast scope in the research area.

References

[1] Abdullah Balamash; Marwan Krunz. "A Client side WWW prefetching model". Proceedings IEEE Global Telecommunications Conference, 2004, Dallas.

[2] Alexander P. Pons. "Improving the performance of client web object retrieval", Journal of Systems and Software, Vol. 74, Issue 3, 2005.

[3] Alexander P. Pons. "Object Prefetching using Semantic Links", Journal of ACM SIGMIS Database, Vol. 37, Issue 1, 2006.

[4] Bin Wu; Ajay D. Kshemakalyani. "Objective-Greedy Algorithms for Long-Term Web Prefetching", Proceedings IEEE International Symposium, NCA'04 , U. S. A.

[5] Bin Wu; Ajay D. Kshemakalyani. "Objective optimal algorithm for long term web prefetching", Journal of IEEE Transactions on Computers, Vol. 55, Issue 1, 2006.

[6] Cheng-Zhong Xu; Tamer I. Ibrahim. "Towards Semantics-Based Prefetching to Reduce Web Access Latency", Proceedings IEEE Computer Society Symposium, SAINT'03, U.S.A.

[7] Chung Ming Huang; TZ Heng Hsu. "A User Aware Prefetching mechanism for video streaming", ACM Journal World Wide Web, Vol. 6, Issue 4, 2003.

[8] Cheng- Zhong Xu; Tamer I. Ibrahim. "Semantics-Based Personalized Prefetching to Improve Web performance",

Proceedings IEEE Conference on Distributed Computing System, U.S.A.
[9] Cheng-Zhong Xu; Tamer I. Ibrahim. "A Keyword-Based Semantic Prefetching Approach in Internet News Services", Journal of IEEE Transactions on Knowledge and Data Engineering, Vol. 16, Issue5, 2004.
[10] Doug Joseph; Dirk Grunwald. "Prefetching Using Markov Predictors, Journal of IEEE Transactions on Computers, Vol. 48, Issue.2, 1999.
[11] George Pallis; Athena Vakali, Jaroslav Pokorny(. "A clustering-based prefetching scheme on a Web cache environment", ACM Journal Computers and Electrical Engineering, Vol.34, Issue 4, 2008.
[12] Holger Kirchner; Reto Krummenacher; David Edwards and Thomas Rissel. "A Location-aware Prefetching Mechanism", Project work at Distributed Information Systems Laboratory LSIR, 2004.
[13] Lei Shi1; Yingjie Han; Xiaoguang Ding1; Lin Wei; Zhimin Gu1. "SPN model for Web Prefetching and Caching", Proceedings IEEE Conference on Semantics, Knowledge and Grid, 2005,Beijing.
[14] Lei Shi; Ying-Jie Han; Xiao Guang Ding; Lin Wei. "An SPN based Integrated Model for Web Prefetching and Caching", Springer Journal of Computer Science and Technology, Vol. 21, Issue 4,2006.
[15] Nizar R. Mabroukeh; Ezeife, C.I.. "Semantic-rich Markov Models for Web Prefetching", Proceedings IEEE International Conference on Data Mining Workshops, 2009, Florida.
[16] N.J. Tuah; M. Kumar; S. Venkatesh. "Resource-aware Speculative Prefetching in Wireless Networks" , Springer Journal on Wireless Networks, Vol. 9, Issue 1,2003.
[17] Philip A. Bernstein; Shankar Pal; David Shutt. "Context-based prefetch – an optimization for implementing objects on relations", ACM, The VLDB Journal, Vol. 9, Issue 3, 2000.
[18] QIANG YANG; Kowloon. "Integrating Web Prefetching and Caching Using Prediction Models", Springer, Journal on World Wide Web, Vol.4, Issue 4, 2001.
[19] Robert Cooksey. "Content-Sensitive Data Prefetching", A thesis submitted to the University of Colorado requirements for the degree of Doctor of Philosophy, 2002.
[20] Sajid Hussai; Robert D.McLeod. "Intelligent Prefetching at a Proxy Server", Proceedings IEEE Conference on Electrical and Computer Engineering, Cananda.
[21] Sang-Won Kang; Jongwan Kim; SeokJin Im; HaRim Jung; Chong-Sun Hwang. "Cache Strategies for Semantic Prefetching Data", Proceedings IEEE Conference on Web-Age Information Management Workshops, China.
[22] Sang-Won Kang; Joon-Min Gil; Jongwan Kim; Seokim; Sangkeun Lee. "Description-Based Semantic Prefetching Schemefor Data Management in Location-Based Services", Journal of Information Science and Engineering, Vol. 24, Issue 6, 2008.
[23] Sarina Sulaiman; Siti Mariyam Shamsuddin. "Web Caching and Prefetching: What, Why, and How?", Proceedings IEEE International Symposium, Malaysia.
[24] Surendra Byna Yong; Chen Xian-He Sun. "A Taxonomy of Data Prefetching Mechanisms", Proceedings IEEE International Symposium on Parallel Architectures, Algorithms, and Networks.
[25] Swapnil Bhatia; Elizabeth Varki; Arif Merchant. "Sequential prefetch cache sizing for maximal hit rate", Proceedings IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication.
[26] Toufiq Hossain Kazi; Wenying Feng; Gongzhu Hu. "Web Object Prefetching: Approaches and a New Algorithm", Proceedings IEEE International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing, 2010.
[27] V. N. Padmanabhan; J. C. Mogul. "Using Predictive Prefetching to Improve World Wide Web Latency," Proceedings ACM Conference, SIGCOMM, U. S. A.
[28] Wei-Guang Teng; Cheng-Yue Chang; Ming-Syan Chen. "Integrating Web Caching and Web Prefetching in Client-Side Proxies", Journal of IEEE Transactions on Parallel and Distributed Systems, Vol.16, Issue 5, 2005.
[29] Xin Jin; Huanqing Xu. "An Approach to Intelligent Web Pre-fetching Based on Hidden Markov Model", Proceedings IEEE Conference on Decision and Control, USA.
[30] Yong Chen; Huaiyu Zhu; Hui Jin; Xian-He Sun. "Improving the Effectiveness of Context-based Prefetching with Multi-order Analysis", Proceedings IEEE International Conference on Parallel Processing Workshops.
[31] Yi- Hung; Arbee L.P. Chen. "Prediction of Web Page Accesses by Proxy Server Log", ACM Journal of World Wide Web, Vol. 5, Issue 1, 2002.
[32] Zhen He; Alonso Marquez. "Path and cache conscious prefetching (PCCP)", ACM, The VLDB Journal, Vol. 16, Issue 2, 2007.
[33] Zhijie Ban; Zhimin Gu; Yu Jin. "An Online PPM Prediction Model for Web Prefetching", Proceedings ACM, International Workshop on Web Information and Data Management, U.S.A.
[34] Zhijie Ban; Zhimin Gul; Yu Jin. "A PPM Prediction Model Based on Stochastic Gradient Descent for Web Prefetching", Proceedings IEEE International Conference on Advanced Information Networking and Applications.

MuthuramThangaraj, received his post-graduate degree in Computer Science from Alagappa University, Karaikudi, M.Tech. degree in Computer Science from Pondicherry University and Ph.D. degree in Computer Science from Madurai Kamaraj University, Madurai, TN, South India in 2006. He is now the Associate Professor of Computer Science Department at M.K.University. He is an active researcher in Web mining, Semantic Web and Information Retrieval and has published more than 35 papers in Journals and Conference Proceedings.

Meenatchi V. T, received her post graduate degree in Computer Science from Madurai Kamaraj University, Madurai, M.Phil. degree in Computer Science from Madurai Kamaraj University, Madurai. She is now the Lecturer at Thiagarajar College, Madurai. Her current research topics have included Semantic Web and Web Prefetching.