Effectively Delivering XML Information in Periodic Broadcast Environments

Yongrui Qin[1], Quan Z. Sheng[1], and Hua Wang[2]

[1] School of Computer Science
The University of Adelaide, Adelaide, SA 5005, Australia
\{yongrui,qsheng\}@cs.adelaide.edu.au
[2] Department of Mathematics & Computing
University of Southern Queensland, QLD 4350, Australia
hua.wang@usq.edu.au

Abstract. Existing data placement algorithms for wireless data broadcast generally make assumptions that the clients' queries are already known and the distribution of access frequencies of their queries can be obtained a priori. Unfortunately, these assumptions are not realistic in most real life applications because new mobile clients may join in anytime and clients may be reluctant to disclose their queries (due to privacy concerns). In this work, we study the data placement problem of periodic XML data broadcast in which similar assumptions can be avoided. This is an important issue, particularly when XML becomes prevalent in today's mobile computing devices. Taking advantage of the structured characteristics of XML data, we are able to generate effective broadcast programs based purely on XML data on the server without any knowledge of the clients' access patterns. This not only makes our work distinguished from previous studies, but also enables it to have broader applicability. We present a theoretical analysis of the problem and discuss structural sharing in XML data which forms the basis of our novel greedy data placement algorithm. Finally, we evaluate the proposed placement algorithm through a set of experiments and the results show that our algorithm can effectively place XML data on air and significantly improve the overall access efficiency.

Key words: periodic data broadcast, XML, multi-item Query, data placement algorithm

1 Introduction

Broadcast is one of the basic ways of information access via wireless technologies. In a wireless data broadcast system, the server broadcasts public information to all mobile devices within its transmission range via a downlink broadcast channel. Mobile clients "listen" to the downlink channel and access information of their interest directly when related information arrives. Broadcast is bandwidth efficient because all mobile clients can share the same downlink channel and re-

trieve data from it simultaneously. Broadcast is also energy efficient at the client ends because downloading data costs much less energy than sending data [29].

There are two typical data broadcast modes: (i) *Periodic Broadcast Mode* and (ii) *On-Demand Broadcast Mode* [29]. In the periodic broadcast mode, data are periodically broadcasted on a downlink channel via which the server sends data to clients. Clients only "listen" to that channel and download data they are interested in. In the on-demand broadcast mode, clients send their queries to the server via an uplink channel and then the server considers all submitted requests and decide the contents of next broadcast cycle. In this work, we focus on the periodic broadcast mode since it has many benefits such as saving uplink bandwidth and power at the client ends by avoiding uplink transmissions and effectively delivering information to an unlimited number of clients simultaneously.

The research of XML data broadcast is of great importance and has been attracting more and more research interests [7, 19, 26, 28, 24]. Information expressed in semi-structured formats is widespread over the past years. XML has rapidly gained popularity as a de facto standard to represent semi-structured information in today's Web. For example, delivering information in XML format is popular in Web services and in different kinds of Publish/Subscribe systems. Similarly, broadcasting information in XML format in a wireless environment is also a preferable way due to the prosperity of XML. We will demonstrate a potential application of using XML data in a broadcast environment by detailing a real life scenario in Section 3.

Data placement algorithms determine what data items to be broadcasted by the server and the order of data items on wireless channels, aiming to reduce average waiting time for mobile clients. To a large extent, the data placement problem of XML data is similar to that in multi-item contexts [27, 4] where mobile clients may request multiple items each time. However, there are drawbacks of existing data placement approaches in traditional data broadcast.

Firstly, previous work on multi-item placement problems generally makes assumptions that the clients' queries are already known and the distribution of access frequencies of these queries can be obtained in advance [1, 2, 6, 15, 27, 4]. For example, it is proposed to allow the clients to provide a profile of their interests to the servers [1, 2], but this can lead to privacy concerns. These assumptions significantly limit the practicability of those proposed placement algorithms in real situations as it is difficult to obtain such kind of information before the organization of data on air starts. Some possible reasons include: (i) new mobile clients may join in the network at anytime; and (ii) mobile users may be reluctant to disclose their queries to the server via uplink channel due to expensive communication cost and privacy concerns.

Secondly, in traditional data broadcast systems, appropriate placement can hardly be generated based only on information of data items themselves on the server. Hence, strong assumptions are inevitable for the design of data placement algorithms. Alternatively, some work applies data mining techniques to discover association rules from the history access patterns of a set of data [3]. This avoids

to obtain access patterns of mobile clients on-the-fly. However, the availability of such history access patterns of mobile clients is a necessity.

By contrast, in XML data broadcast, data items (or XML documents) usually share parts of their structure. Taking structural sharing between XML documents into consideration, we are able to analyze and estimate clients' access patterns via the analysis of this structural sharing. Then we can effectively place XML data on wireless channels based purely on XML data on the server, which is important for practical usage. To the best of our knowledge, little work has addressed similar data placement strategies in the context of wireless data broadcast.

In this paper, we study the data placement problem of periodic XML data broadcast. Firstly, we describe the overall system model and present a theoretical analysis on the data placement problem of the periodic XML data broadcast. Secondly, based on the analysis, we design a novel greedy data placement algorithm. In summary, the main contributions of this paper can be described as follows:

- We found that the assumptions on the clients' queries and their distributions that have been made by previous work can be removed in the context of periodic XML data broadcast. By taking advantage of the structural characteristics of XML data, we are able to generate appropriate data placement results based only on XML data on the server.
- We present a theoretical analysis on the data placement problem of periodic XML data broadcast. Based on the analysis, a novel greedy data placement algorithm which organizes XML data on air is presented.
- Extensive experiments are conducted to show the effectiveness of our proposed data placement algorithm.

The remainder of this paper is organized as follows. Section 2 describes background knowledge of this work, including an application scenario, the system model and XML similarity background. Then a theoretical analysis of data placement problem is presented in Section 3. Section 4 discusses the structural sharing property of XML data and proposes a novel greedy data placement algorithm. Section 5 presents our experimental study for evaluating the performance of the proposed data placement algorithm. Finally, Section 6 discusses related work and Section 7 gives some concluding remarks.

2 Application Scenario, System Model and XML Similarity

In this section, we first describe an application scenario. Then we show the system model of this work and introduce background knowledge of XML similarity.

2.1 Application Scenario

We use the following scenario to show potential applications of XML data broadcast in real life.

Consider a live basketball game. Information about the game and the players on the court is usually the interest of a large number of audience. In this context, data broadcast is a preferable way of delivering latest information to the audience. Meanwhile, some audience could be outside of the stadium, such as basketball fans who are watching live text information about the game via the Internet at their homes. Therefore, the game information could also be delivered via the Internet to online audience and other Web service providers who have subscribed this basketball game. Using XML format to represent game information can satisfy all these needs and realize simplicity, generality, and usability of game information at the same time.

2.2 Periodic XML Data Broadcast System Model

Fig. 1 shows the model of our wireless XML data broadcast system. The system includes an XML Data Center (the broadcast server), a broadcast program scheduler, broadcast listeners (mobile clients) and a downlink channel (the server sends information to mobile clients via it). The downlink channel can be shared by all mobile clients. But mobile clients can not send their individual queries to the server in this model as no uplink channel is available.

From the figure, we can see that the XML Data Center could be connected to the Internet and deliver information to online users, Web service providers and Publish/Subscribe systems, etc. With the use of XML format data, these different applications can be integrated seamlessly with our wireless XML data broadcast system for the purpose of sharing and delivering same information to different kinds of users.

2.3 XML Similarity

Some existing work on measuring structural similarity between XML documents can be found in [22, 11]. The main idea of their work is based on the concept of *path sets*. Here, a path set of an XML document contains all full paths (paths that are from root element to leaves) and their subpaths. A simple example is presented in Fig. 2. The path set of this example is: {`/player/name, /player/position, /player/nationality, /player/college, /player, /name, /position, /nationality, /college`}. We denote a path set of an XML document d as $PS(d)$.

Different types of measure can be adopted, such as Jaccard measure [17, 10], Dice's coefficient [9] and Lian's measure [16], to measure the similarity between two XML documents d_i and d_j. The exact forms of these measures based on PS are as follows (Jaccard measure denoted as $J(d_i, d_j)$, Dice's coefficient denoted as $D(d_i, d_j)$ and Lian's measure denoted as $L(d_i, d_j)$):

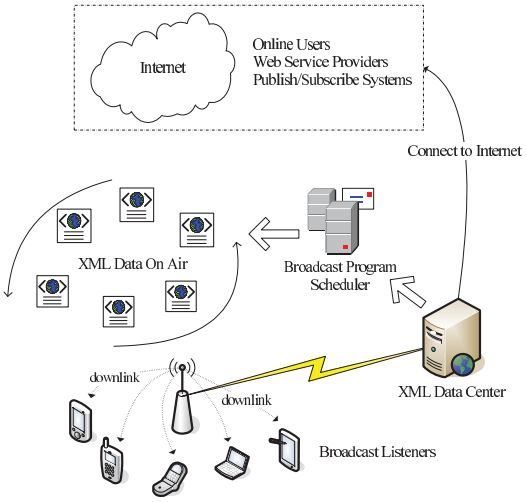


Fig. 1. A wireless XML data broadcast system

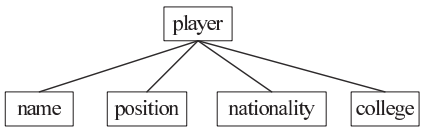


Fig. 2. An XML structure tree

$$J(d_i, d_j) = \frac{|PS(d_i) \bigcap PS(d_j)|}{|PS(d_i) \bigcup PS(d_j)|} \tag{1}$$

$$D(d_i, d_j) = \frac{2 \cdot |PS(d_i) \bigcap PS(d_j)|}{|PS(d_i)| + |PS(d_j)|} \tag{2}$$

$$L(d_i, d_j) = \frac{|PS(d_i) \bigcap PS(d_j)|}{max\{|PS(d_i)|, |PS(d_j)|\}} \tag{3}$$

From the above definitions, we can see that both Jaccard measure and Dice's coefficient give more weights on the total structural information of two comparing documents while Lian's measure emphasizes more on the difference of these documents. All three measures can vary in interval $[0, 1]$. If $PS(d_i) = PS(d_j)$, we have $J(d_i, d_j) = D(d_i, d_j) = L(d_i, d_j) = 1$. Clearly, the larger the values of these measures are, the more structural sharing the two comparing XML documents have.

In the literature, two critical metrics, namely *access time* and *tuning time*, are used to measure the system's performance [12]. Data placement mainly affects access time because tuning time depends on the total content downloaded by mobile clients but not on the order of data. Hence, we use access time as our metric in this analysis. In periodic broadcast, queries are used to describe the interests of mobile clients and help mobile clients to skip irrelevant data on air, but they are not actually submitted to the broadcast server.

3 Analysis of Data Placement Problem

In this section, we present a theoretical analysis on the data placement problem in periodic XML data broadcast.

In the literature, two critical metrics, namely *access time* and *tuning time*, are used to measure the system's performance [12]. Data placement mainly affects access time because tuning time depends on the total content downloaded by mobile clients but not on the order of data. Hence, we use access time as our metric in this analysis. In periodic broadcast, queries are used to describe the interests of mobile clients and help mobile clients to skip irrelevant data on air, but they are not actually submitted to the broadcast server.

Table 1 lists the symbols used in the rest of the paper and Fig. 3 shows a broadcast program (or broadcast sequence) σ on the wireless channel which is broadcasted periodically. The broadcast program σ can start from any XML document d_i. However, we assume that σ starts from d_1 (this will then comply with the definition of σ in Table 1) to simplify our analysis.

With the basic assumption that queries can be issued at any time with an equal probability (this means the issue time of queries follows uniform distribution), we can calculate the expected access time of q, denoted as $\mathcal{AT}^q_{exp}$, in the following:

$$\begin{aligned}\mathcal{AT}^q_{exp} &= \sum_{i=1}^{k}\left(\frac{\mathcal{L}_{d_{n_i}}}{\mathcal{L}_\sigma}\cdot\mathcal{L}_\sigma + \frac{\mathcal{L}_{gap_i}}{\mathcal{L}_\sigma}\cdot\left(\mathcal{L}_\sigma - \frac{1}{2}\cdot\mathcal{L}_{gap_i}\right)\right)\\ &= \sum_{i=1}^{k}\mathcal{L}_{d_{n_i}} + \sum_{i=1}^{k}\mathcal{L}_{gap_i} - \frac{1}{\mathcal{L}_\sigma}\cdot\sum_{i=1}^{k}\frac{1}{2}\cdot\mathcal{L}^2_{gap_i}\\ &= \mathcal{L}_\sigma - \frac{1}{2\cdot\mathcal{L}_\sigma}\cdot\sum_{i=1}^{k}\mathcal{L}^2_{gap_i} \qquad (4)\end{aligned}$$

According to Equation (4)[1] and a given broadcast program σ, we can calculate $\mathcal{AT}^q_{exp}$ simply according to the gaps between consecutive documents required by q. Further, from the above equation, we can see that in order to improve expected access efficiency, $\sum_{i=1}^{k}\mathcal{L}^2_{gap_i}$ should be as large as possible.

[1] This result is exactly the same as [6] although the deduction process is different. The further analysis on this result in the following is new.

Table 1. Symbols Overview

Symbol	Description
$\mathcal{D}$	XML document set. $\{d_1, d_2, d_3, \ldots, d_n\}$.
n	number of XML documents in $\mathcal{D}$.
q	query issued by a mobile client.
$\mathcal{D}_q$	a set of XML documents required by q. $\{d_{n_1}, d_{n_2}, d_{n_3}, \ldots, d_{n_k}\}$.
k	number of documents required by q.
σ	a complete broadcast program (or broadcast sequence). It is the result of a data placement algorithm running on a given $\mathcal{D}$. $< d_1, d_2, d_3, \ldots, d_n >$. (Note: It is different from $\mathcal{D}$ as $\mathcal{D}$ is a set but not a sequence.)
σ_q	a subsequence of σ which includes only k documents required by q. $< d_{n_1}, d_{n_2}, d_{n_3}, \ldots, d_{n_k} >$. (Note: It is different from $\mathcal{D}_q$ as $\mathcal{D}_q$ is a set but not a sequence.)
gap_i	unmatched documents of q that are placed between d_{n_i} and $d_{n_{i+1}}$ in σ. Note that d_{n_i} and $d_{n_{i+1}}$ are consecutive documents in $\sigma_q (1 \leq i < k)$.
gap_k	unmatched documents of q that are placed between d_{n_k} and d_{n_1} in two consecutive σ (Note: σ will be broadcasted periodically).
$\mathcal{L}_{gap_i}$	the total length of all unmatched documents in gap_i $(1 \leq i \leq k)$.
$\mathcal{L}_{gaps}$	the total length of all gaps, which is $\sum_{i=1}^{k} \mathcal{L}_{gap_i}$.
$\mathcal{L}_{d_i}$	the length of an XML document d_i.
$\mathcal{L}_{d_{n_i}}$	the length of an XML document d_{n_i} which is the i^{th} document in σ_q.
$\mathcal{L}_{\sigma}$	the length of σ, which is $\sum_{i=1}^{n} \mathcal{L}_{d_i}$.
$\mathcal{L}_{\sigma_q}$	the length of σ_q, which is $\sum_{i=1}^{k} \mathcal{L}_{d_{n_i}}$.
$\mathcal{AT}_{exp}^{q}$	the expected access time of q.

Moreover, according to definitions in Table 1, the sum of all gaps, denoted $\mathcal{L}_{gaps}$, can be computed as

$$\mathcal{L}_{gaps} = \sum_{i=1}^{k} \mathcal{L}_{gap_i} = \mathcal{L}_{\sigma} - \sum_{i=1}^{k} \mathcal{L}_{d_{n_i}} = \mathcal{L}_{\sigma} - \mathcal{L}_{\sigma_q} \tag{5}$$

Note that $\mathcal{L}_{\sigma_q}$ is independent of any data placement results. In other words, $\mathcal{L}_{\sigma_q}$ is fixed for a given q, which in turn indicates that $\mathcal{L}_{gaps}$ is fixed.

In order to derive lower and upper bounds of $\sum_{i=1}^{k} \mathcal{L}_{gap_i}^2$ and to analyze our data placement strategy, we first present the following propositions for below function $f(\mathbf{X})$.

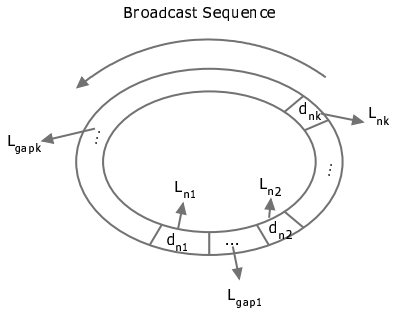


Fig. 3. A broadcast program showing positions of documents required by query q

In order to derive lower and upper bounds of $\sum_{i=1}^{k} \mathcal{L}_{gap_i}^2$ and to analyze our data placement strategy, we first present the following propositions for below function $f(\mathbf{X})$.

Function $f(\mathbf{X}) = {x_1}^2 + {x_2}^2 + \ldots + {x_k}^2$ is with the following constraints:

1. $x_1 + x_2 + \ldots + x_k = M$
2. $x_1, x_2, \ldots, x_k \geq 0$

where M is a positive constant. We also denote the lower bound and the upper bound of $f(\mathbf{X})$ as $\underline{f(\mathbf{X})}$ and $\overline{f(\mathbf{X})}$ respectively.

Proposition 1 *Given $f(\mathbf{X})$ defined as above, we must have*

$$f(\mathbf{X}) \leq M^2$$

When $x_1 = x_2 = \ldots = x_{k-1} = 0$ and $x_k = M$ (or any other kind of combinations like this), $f(\mathbf{X})$ reaches its upper bound, i.e., $\overline{f(\mathbf{X})} = M^2$.

Proposition 2 *Given $f(\mathbf{X})$ defined as above, we must have*

$$f(\mathbf{X}) \geq k \cdot (\frac{M}{k})^2$$

When $x_1 = x_2 = \ldots + x_k = \frac{M}{k}$, $f(\mathbf{X})$ reaches its lower bound, i.e., $\underline{f(\mathbf{X})} = k \cdot (\frac{M}{k})^2$.

Moreover, given $f(\mathbf{X})$ defined as above and suppose that m variables, i.e. $x_1, x_2, \ldots, x_m$, have been determined ($m < k$) while the rest $k - m$ variables are not. We also denote $M' = M - \sum_{i=1}^{m} x_i$. Now we are going to determine next variable. Without loss of generality, we use x_{m+1} as our next variable to be determined and aim to maximize or minimize $f(\mathbf{X})$. We denote $f(\mathbf{X})_{x_{m+1}}$ as

the function with $m+1$ determined variables (x_i, $1 \leq i \leq m+1$) and $k-m-1$ undetermined variables. Then given two values of this variable, i.e. x_{m+1} and x'_{m+1} and suppose $x_{m+1} < x'_{m+1}$. Then we have the following propositions.

Proposition 3 *For* $\overline{f(\mathbf{X})}_{x_{m+1}}$ *and* $\overline{f(\mathbf{X})}_{x'_{m+1}}$, *we have*

$$\begin{cases} \overline{f(\mathbf{X})}_{x_{m+1}} > \overline{f(\mathbf{X})}_{x'_{m+1}} & x_{m+1} < x'_{m+1} \leq \frac{M'}{2} \\ \overline{f(\mathbf{X})}_{x_{m+1}} < \overline{f(\mathbf{X})}_{x'_{m+1}} & \frac{M'}{2} \leq x_{m+1} < x'_{m+1} \\ Indefinite & Otherwise \end{cases}$$

Proposition 4 *For* $\underline{f(\mathbf{X})}_{x_{m+1}}$ *and* $\underline{f(\mathbf{X})}_{x'_{m+1}}$, *we have*

$$\begin{cases} \underline{f(\mathbf{X})}_{x_{m+1}} > \underline{f(\mathbf{X})}_{x'_{m+1}} & x_{m+1} < x'_{m+1} \leq \frac{M'}{k-m} \\ \underline{f(\mathbf{X})}_{x_{m+1}} < \underline{f(\mathbf{X})}_{x'_{m+1}} & \frac{M'}{k-m} \leq x_{m+1} < x'_{m+1} \\ Indefinite & Otherwise \end{cases}$$

The proofs of these propositions can be found in the Appendix. Now according to Proposition 1 and Proposition 2, we have

$$k \cdot (\frac{\mathcal{L}_{gaps}}{k})^2 \leq \sum_{i=1}^{k} \mathcal{L}_{gap_i}^2 \leq \mathcal{L}_{gaps}^2 \tag{6}$$

Then according to Equation (4), we have

$$\mathcal{L}_\sigma - \frac{\mathcal{L}_{gaps}^2}{2 \cdot \mathcal{L}_\sigma} \leq \mathcal{AT}_{exp}^q \leq \mathcal{L}_\sigma - \frac{\mathcal{L}_{gaps}^2}{2 \cdot k \cdot \mathcal{L}_\sigma} \tag{7}$$

From the above two inequations, we can see that in order to improve expected access efficiency, $\sum_{i=1}^{k} \mathcal{L}_{gap_i}^2$ should be as large as possible. According to Proposition 1, when we have one of the gaps equal to $\mathcal{L}_{gaps}$ and all other gaps equal to 0, we can achieve best expected access efficiency. Thus, when all XML documents required by q are placed together and broadcasted in sequence, $\mathcal{AT}_{exp}^q$ can be minimized. Also, according to Equation (5), we can rewrite the above inequation to

$$\mathcal{L}_\sigma - \frac{(\mathcal{L}_\sigma - \mathcal{L}_{\sigma_q})^2}{2 \cdot \mathcal{L}_\sigma} \leq \mathcal{AT}_{exp}^q \leq \mathcal{L}_\sigma - \frac{(\mathcal{L}_\sigma - \mathcal{L}_{\sigma_q})^2}{2 \cdot k \cdot \mathcal{L}_\sigma} \tag{8}$$

Here Inequation (8) shows both the lower and upper bounds of $\mathcal{AT}^q_{exp}$ for q in another form. It is worth mentioning that both bounds are independent of any data placement results. Moreover, we can infer when k increases, σ_q will include more documents. Then $\mathcal{L}_{\sigma_q}$ increases as well. However, the decrease of difference $\mathcal{L}_\sigma - \mathcal{L}_{\sigma_q}$ leads to larger lower and upper bounds of $\mathcal{AT}^q_{exp}$, which means the system's overall performance will degrade.

The above analysis focuses on a single query. However, generalizing it to multiple queries would be much more complicated. Actually, determining an optimal broadcast sequence for multiple multi-item queries is an NP-Complete problem [6].

When there are multiple queries to consider for a broadcast program, these queries are not likely to require the same XML documents. In such cases, Proposition 1 and Inequation (6), which minimizes expected access time for a single query, cannot help to find an optimal solution for all queries. But according to Proposition 3 and Proposition 4, we know that when $x_{m+1} \leq \frac{M'}{2}$ and $x_{m+1} \leq \frac{M'}{k-m}$, we should decrease x_{m+1} to have larger $\overline{f(\mathbf{X})}_{x_{m+1}}$ and $\underline{f(\mathbf{X})}_{x_{m+1}}$. In other words, if we progressively reduce each gap as much as possible, we would have larger lower and upper bounds of $\sum_{i=1}^{k} \mathcal{L}^2_{gap_i}$. In this way, we can reduce both the lower and upper bounds of overall expected access time.

For example, if we need to minimize $\sum_{i=1}^{k} \mathcal{L}^2_{gap_i}$ for each query in $\{q_1, q_2, q_3\}$, for the first step, we should place XML documents that are required by all three queries together to form an initial broadcast program. In the second step, we should place XML documents required by two of the three queries together and append them to the initial broadcast program. After that, we append XML documents required by only one query to the broadcast program to form a final broadcast program. In this way, we can construct a final broadcast program in a greedy style.

Now the problem becomes how we can determine which documents should be placed together first as we cannot obtain queries in advance. Our solution will be discussed in next section.

4 Data Placement Algorithm

In this section, we introduce our data placement algorithm for periodic XML data broadcast. This algorithm is based on the theoretical analysis in previous section. We first discuss the structural sharing property of XML data which we use to estimate the potential access patterns of mobile clients, i.e., the probability of accessing a small set of similar XML documents simultaneously. Then we put forward a novel greedy data placement algorithm based on it.

4.1 Structural Sharing in XML data

Intuitively, for any two given XML documents, we can utilize one of the three similarity measures described in Section 2.3 to calculate the similarity between

them and the similarity results can be used to approximate the probability that a specific query is matched with both documents at the same time. For example, if two XML elements are under structurally similar paths, then it is more likely that either both elements or none satisfy a given query [22]. In fact, query issuers hardly have thorough knowledge about the broadcasted content and XPath queries usually contain * and // which would match similar structure. Therefore, if two XML documents are with larger structural similarity, i.e. d_1 and d_2, then they would have a higher probability to be required simultaneously. However, there are still three other cases to be considered, such as required d_1 but not d_2, required d_2 but not d_1 and required neither of d_1 and d_2. Therefore, the above similarity measures consider only successful match probabilities of both XML documents but do not consider unsuccessful match probabilities of them.

Nonetheless, unsuccessful match cases have effects on the expected access time as well. According to Proposition 1 and Proposition 2, in order to have better access efficiency, the gaps between any two required documents by a single query should be as less uniform as possible. Based on this, we can infer that in the above example, cases of required d_1 but not d_2 and required d_2 but not d_1 are likely to generate more uniform gaps while other two cases (required both documents or neither) are likely to have less uniform gaps. Observing this, we define a new similarity measure called *Cohesion* to give a more accurate estimation of access patterns of mobile clients in the following.

Note that, for any query q requiring at least one of the documents in $\mathcal{D}$, q must match some paths in $PS(\mathcal{D})$ and it has a probability of $\frac{|PS(d)|}{|PS(\mathcal{D})|}$ to match d. If a query q fails to match any document in $\mathcal{D}$, the issuer of q only needs to locate and download air index to confirm that his/her query does not match any document. Then he/she can stop waiting the result to be broadcasted. All such kind of queries only need to access the index information on air and therefore, their expected access time depends heavily on the index distribution, which is not the focus of our work. To estimate their expected access time, interested readers are referred to [12] for more details. Hence, we only consider successful queries in this work.

Now suppose we have a set of n XML documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ on the server, we can approximate access probability of any document d for queries which successfully match at least one document in set $\mathcal{D}$ as follows:

$$Pr(d) = \frac{|PS(d)|}{|PS(\mathcal{D})|} \tag{9}$$

and for any i, j $(1 \leq i, j \leq n)$

$$Pr(d_i - d_j) = \frac{|PS(d_i) - PS(d_j)|}{|PS(\mathcal{D})|} \tag{10}$$

Here, $PS(\mathcal{D}) = \bigcup_{i=1}^{n} PS(d_i)$.

There would be many different matching cases for a given set $\mathcal{D}$. Take two XML documents d_1 and d_2 in $\mathcal{D}$ as an example. As mentioned previously, there

Table 2. Matching Cases for Document d_1 and d_2 in a document set $\mathcal{D}$

Case	Probability	Effect on AT_{exp}
Matched both d_1, d_2	$Pr(d_1 \bigcap d_2)$	Positive
Matched none of d_1, d_2	$1 - Pr(d_1 \bigcup d_2)$	Positive
Matched d_1, but not d_2	$Pr(d_1 - d_2)$	Negative
Matched d_2, but not d_1	$Pr(d_2 - d_1)$	Negative

would be four cases of matching of them and the probability of each case is shown in Table 2. In this table, we also include positive and negative effects on the expected access time (AT_{exp}) for each case.

Based on Table 2, we define Cohesion $C(d_i, d_j)$ of XML documents d_i and d_j as follows:

$$C(d_i, d_j) = \frac{Pr(d_i \bigcap d_j) \cdot (1 - Pr(d_i \bigcup d_j))}{max\{Pr(d_i - d_j), Pr(d_j - d_i)\}} \tag{11}$$

Here d_i and d_j are both in set $\mathcal{D}$. It is easy to see that $C(d_i, d_j) = C(d_j, d_i)$. According to Equation (9), Equation (10) and Equation (11), we can calculate $C(d_i, d_j)$ after finding path sets of d_i, d_j and $\mathcal{D}$. Cohesion values can vary in a wide range which exceeds interval $[0, 1]$. Strictly speaking, Cohesion values only vary in interval $[0, \frac{|PS(\mathcal{D})|}{4}]$ given that $C(d_i, d_j) = \frac{|PS(\mathcal{D})|}{4}$ when $PS(d_i) = PS(d_j)$. The lower bound 0 is trivial. In order to obtain the upper bound, we only consider cases that have $PS(d_i) \neq PS(d_j)$, from which we can infer that $max\{|PS(d_i - d_j)|, |PS(d_j - d_i)|\} \geq 1$. Without loss of generality, let $|PS(d_i)| \geq |PS(d_j)|$, according to Equation (9) and Equation (10), we can rewrite Equation (11) as follows:

$$\begin{aligned} C(d_i, d_j) &\leq \frac{\frac{|PS(d_i \bigcap d_j)|}{|PS(\mathcal{D})|} \cdot (1 - \frac{|PS(d_i \bigcup d_j)|}{|PS(\mathcal{D})|})}{\frac{1}{|PS(\mathcal{D})|}} \\ &< \frac{|PS(d_i)| \cdot (|PS(\mathcal{D})| - |PS(d_i)|)}{|PS(\mathcal{D})|} \\ &= \frac{-(|PS(d_i)| - \frac{|PS(\mathcal{D})|}{2})^2 + \frac{|PS(D)|^2}{4}}{|PS(\mathcal{D})|} \\ &\leq \frac{|PS(\mathcal{D})|}{4} \end{aligned}$$

Then the above result gives the upper bound of Cohesion $C(d_i, d_j)$. Now we can normalize Cohesion values to interval $[0, 1]$ in the following

$$C'(d_i, d_j) = \begin{cases} \frac{4 \cdot C(d_i, d_j)}{|PS(\mathcal{D})|} & PS(d_i) \neq PS(d_j) \\ 1 & PS(d_i) = PS(d_j) \end{cases} \tag{12}$$

Algorithm 1 Initialize structural sharing matrix $S[n][n]$

Input: A set of XML documents $\mathcal{D} : \{d_1, d_2, ..., d_n\}$
Output: Structural sharing matrix $S[n][n]$
1. create matrix $S[n][n]$
2. **for** each document d in $\mathcal{D}$ **do**
3. compute $PS(d)$
4. **end for**
5. **for** each pair of documents $< d_i, d_j >$ in $\mathcal{D}$ $(i < j)$ **do**
6. $S[i][j] \Leftarrow$ structural sharing between d_i and d_j
7. $S[j][i] \Leftarrow S[i][j]$
8. **end for**

We can also infer that $C'(d_i, d_j) = 1$ if and only if $PS(d_i) = PS(d_j)$. Similar to other three similarity measures, the larger the value of Cohesion is, the more structural sharing the two comparing XML documents have.

Algorithm 2 GDPA

Input: Structural sharing matrix $S[n][n]$
Output: A broadcast program σ for $\mathcal{D}$
1. $\sigma \Leftarrow$ empty sequence
2. select a pair of documents $< d_i, d_j >$ with maximum value $S[i][j]$ in matrix $S[n][n]$
3. **if** $L_{d_1} <= L_{d_j}$ **then**
4. add $< d_i, d_j >$ into σ
5. **else**
6. add $< d_j, d_i >$ into σ
7. **end if**
8. $\mathcal{D}' \Leftarrow \mathcal{D} - d_i - d_j$
9. **while** $\mathcal{D}'$ is not empty **do**
10. $d_{head} \Leftarrow$ the first document in σ
11. select a pair of documents $< d_{i_{max}}, d_{head} >$ with maximum value $S[i_{max}][head]$ $(d_{i_{max}} \in \mathcal{D}')$
12. $d_{rear} \Leftarrow$ the last document in σ
13. select a pair of documents $< d_{j_{max}}, d_{rear} >$ with maximum value $S[j_{max}][rear]$ $(d_{j_{max}} \in \mathcal{D}')$
14. **if** $S[i_{max}][head] >= S[j_{max}][rear]$ **then**
15. append $d_{i_{max}}$ into σ from head
16. $\mathcal{D}' \Leftarrow \mathcal{D}' - d_{i_{max}}$
17. **else**
18. append $d_{i_{max}}$ into σ from rear
19. $\mathcal{D}' \Leftarrow \mathcal{D}' - d_{j_{max}}$
20. **end if**
21. **end while**

4.2 The Greedy Data Placement Algorithm

Based on the discussion of structural sharing in XML data, we can generate a broadcast program for periodic data broadcast in a greedy way. From previous discussions, we can see that the more the structural sharing of two XML documents is, the larger probability of matching both XML documents simultaneously. As a result, our Greedy Data Placement Algorithm (GDPA) places XML documents with most structural sharing together first as an initial broadcast program. Then it progressively appends other XML documents to the broadcast program in a descendant order of structural sharing. Detailed steps of GDPA are shown in Algorithm 1 and Algorithm 2.

Algorithm 1 initializes a structural sharing matrix $S[n][n]$ for n XML documents on the broadcast server. Note that, all four similarity measures defined in subsection 2.3 and 4.1 can be used in Algorithm 1 to compute structural sharing between two documents (Line 6). All of them are symmetric which means for any one of these measures, we must have $S[j][i] = S[i][j]$. Also we have $J(d_i, d_j) = D(d_i, d_j) = L(d_i, d_j) = C'(d_i, d_j) = 1$ if $i = j$. Therefore, we only need to calculate matrix S for entries $S[i][j]$ where $i < j$.

Based on matrix S, Algorithm 2 finds the pair of XML documents with maximum structural sharing value and adds them into the initial empty broadcast program σ (Line 2). As discussed in Section 3, the expected access time is determined by the gaps between the required documents but not by the sequence of them. Therefore, the sequence of the first pair of XML documents can be simply placed according to the ascendant order of document lengths (Line 3 to 7). Then Algorithm 2 appends the XML document with maximum structural sharing to the head document d_{head} or the rear document d_{rear} of σ. If the maximum structural sharing is derived between document d and document d_{head}, d will be appended into σ from head; otherwise, d will be appended into σ from rear. This process will be repeated until all XML documents are placed into σ in order (Line 9 to 21).

5 Experiments

In this section, we study the performance of our data placement algorithm. We show its efficiency in terms of access time, which is a common measure of performance in data broadcasts. Since this is the first work that determines broadcast schedules based only on XML data on the server without any knowledge of the clients' access patterns, we compare our algorithm with a common random data placement algorithm (RDPA).

5.1 Experimental Setup

The experiments are run on three data sets each with 250 XML documents defined by News Industry Text Format (NITF) DTD [13], which is published for news copy production, press releases, and Web-based news organizations. The

average depth of the three document sets is between 6 and 8 while the maximum depth is 20.

The details of these data sets are shown in Table 3. Data in $DS1$ can be well clustered into 6 clusters. Moreover, for any two documents d_i, d_j in two different clusters of $DS1$, the minimum similarity values, the maximum similarity values and the average similarity values of all four measures (normalized Cohesion is adopted here) are shown in Table 4. We can see that all clusters are quite different from each other and share very little structural information. Data in $DS2$ are miscellaneous. Documents in $DS2$ cannot be classified into fine clusters. Data in $DS3$ are a mix of well-clustered data and miscellaneous data, which include 125 XML documents from $DS1$ and 125 XML documents from $DS2$.

Table 3. Data Sets in Our Experiments

Set Name	Length			Remark
	Minimum	Maximum	Average	
$DS1$	2.4KB	8.1KB	5.0KB	6 clusters
$DS2$	0.5KB	45.9KB	12.4KB	miscellaneous
$DS3$	2.4KB	24.8KB	9.9KB	hybrid

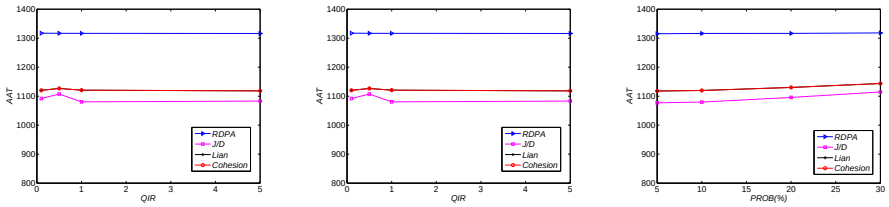
Table 4. Similarity between clusters in $DS1$

Measure	Similarity		
	Minimum	Maximum	Average
Jaccard	0.0097	0.1102	0.0435
Dice	0.0049	0.0583	0.0225
Lian	0.0057	0.1039	0.0345
Cohesion	0.0229	0.4620	0.1457

Table 5. Workload Parameters for the Experiments

Parameter	Range	Default	Description
$PROB$	5% to 30%	10%	probability(* and //)
QIR	0.1 to 5	1	query incoming rate
MQD	5 to 8	7	maximum query depth

In the experiments, XPath queries are generated using the generator developed by [8]. Queries are allowed to repeat. The generator provides several parameters to generate different types of XPath queries, such as query depth, probability of * and // and so on. The probability of * and // appearing in each



(a) Varying $PROB$ (b) Varying QIR (c) Varying MQD

Fig. 4. Evaluating AAT Performance on $DS1$: well-clustered data set

query's step is between 5% and 30% (denoted $PROB$, and the default value is 10%). Query Incoming Rate (denoted QIR) means the number of newly issued queries from mobile clients in a unit of time. We measure this unit of time by the time that mobile wireless system takes to broadcast a block of 1024-byte XML data. The maximum depth of generated XPath queries (denoted MQD) is between 5 and 8. Table 5 shows the value range of parameters in the experiments.

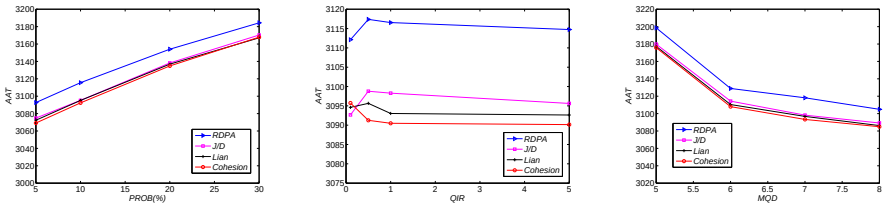
The random data placement algorithm (denoted RDPA) is compared with GDPA (implemented using all four similarity measures defined in Equations (1), (2), (3) and (12)). In RDPA, the server broadcasts XML documents in a random order. This random order is implemented by a Java class `Random`.

We implement both RDPA and GDPA on Java Platform Standard Edition 6 running on Windows 7 Enterprise, 64-bit Operating System. All our experiments are obtained by running 30 consecutive broadcast cycles. When we vary $PROB$, we set QIR and MQD to their default values. When we vary QIR, we set $PROB$ and MQD to their default values. Similarly, when we vary MQD, we set $PROB$ and QIR to their default values.

Regarding air indexing and index distribution strategy, in our experiments, we adopt Compact Index (CI) [26] as our index structure and $(1, m)$ index scheme [12] as our index distribution strategy. This is because CI is the state-of-the-art indexing technique for XML data broadcast and $(1, m)$ index scheme is the most popular index distribution strategy for traditional periodic data broadcast. More details can be found in [26] and [12].

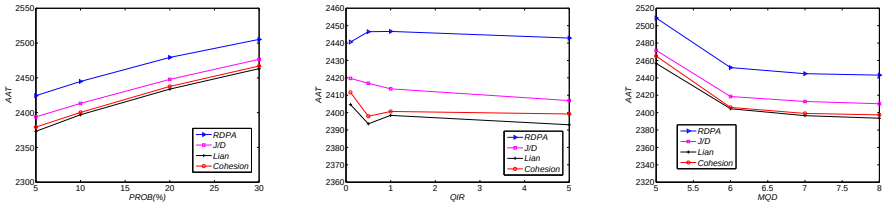
5.2 Performance of GDPA

Our experimental results are shown in Fig. 4, Fig. 5 and Fig. 6. Average access time (AAT) is our performance metric. Also we only consider AAT for all successful matched queries and abandon unsuccessful matched queries. The main reason for this is that, AAT of unsuccessful queries is determined by index distribution but not by data placement results (more details about this can be found in [12]). Note that, GDPA can be implemented with four different similarity measures defined in Section 4, which are Jaccard measure, Dice's coefficient, Lian's measure and our proposed Cohesion. Through our experiments, Jaccard



(a) Varying $PROB$ (b) Varying QIR (c) Varying MQD

Fig. 5. Evaluating AAT Performance on $DS2$: miscellaneous data set



(a) Varying $PROB$ (b) Varying QIR (c) Varying MQD

Fig. 6. Evaluating AAT Performance on $DS3$: a mixed set of well-clustered data and miscellaneous data

measure and Dice's coefficient always yield the same results. Therefore, we denote GDPA implemented with them as *J/D* method in all figures. Meanwhile, we denote GDPA implemented with Lian's measure as *Lian* method and denote GDPA implemented with Cohesion as *Cohesion* method.

Fig. 4 shows the results on $DS1$. From the figure we can see that all GDPA methods outperform RDPA significantly. Specifically, J/D method achieves the best results while Lian method and Cohesion method provides similar results. This indicates that J/D method better fits well-clustered data. In Fig. 4(a), GDPA methods become slightly worse when $PROB$ increases. Since $DS1$ is well-clustered, most queries only require documents in the same clusters. Thus $PROB$ has less effect on AAT. In Fig. 4(b), when QIR increases, J/D method becomes slightly better. This indicates that J/D method can achieve better scalability than other methods when accessing well-clustered data. Fig. 4(c) shows that all GDPA methods remain stable as MQD increases. It is interesting to note that for RDPA, AAT always remains stable.

Fig. 5 shows the results on $DS2$. From the figure we can see that all GDPA methods achieve better performance when compared with RDPA. Specifically, Cohesion method achieves the best results while J/D method achieves the worst results among GDPA methods. This indicates that Cohesion method better fits miscellaneous data. In Fig. 5(a), both GDPA methods and RDPA become worse when $PROB$ increases. It is clear that $PROB$ has more effect on AAT for miscellaneous data. In Fig. 5(b), when QIR increases from 0.1 to 0.5, GDPA methods

J/D and Lian together with RDPA become worse while Cohesion method still becomes better. After that, when QIR increases, all methods become slightly better. This shows that Cohesion method can achieve best scalability when accessing miscellaneous data.

Fig. 6 shows the results on $DS3$. Similarly, all GDPA methods achieve better performance when compared with RDPA. Specifically, Lian method achieves the best results while J/D method provides the worst results among GDPA methods. This shows that Lian method better fits hybrid data. However, Cohesion method achieves very similar performance of Lian method. In Fig. 6(a), both GDPA methods and RDPA become worse when $PROB$ increases. $PROB$ has more effect on AAT for hybrid data. In Fig. 6(b), when QIR increases, all GDPA methods become slightly better and still Lian method provides the best results.

To sum up, GDPA methods always achieve better AAT when compared with RDPA. When accessing well-clustered data, J/D method achieves the best performance. When accessing miscellaneous data, Cohesion method provides the best performance and finally when accessing hybrid data, Lian method shows the best performance.

6 Related Work

Many studies have been done to investigate data placement techniques to reduce access time [25, 23, 14]. These studies generally assume that each user query requires one data item only. Other studies handle data placement problems for queries that may require multiple data items.

Multi-item data placement problem is related to the data placement problem of XML data which is the focus of our work. It is proved to be a NP-Complete problem [6]. A data placement method for multi-item queries called QEM is introduced in [5], which opened up a new perspective in this field. In addition, several improved methods are proposed [15, 3]. The above work is all within the scope of periodic broadcast and generally makes assumptions that the clients' queries are already known and the distribution of access frequencies of these queries can be obtained in advance. However, these assumptions are not true for most applications in real life because the demand is either not known or it may be costly to collect the demand information.

Multi-item data placement problem in on-demand broadcast mode has also attracted lots of interests [27, 21]. These approaches are in pure on-demand broadcast mode and strictly require that mobile clients submit their queries to the server for desired data. Otherwise, the server will not broadcast related data on air. This is because the server filters and schedules data solely based on submitted queries. However, frequent use of uplink channel leads to high communication cost via uplink channel, which can shorten battery life of mobile clients dramatically.

The above mentioned studies focus on flat data broadcasts, in which indices of data items are generally key-based and data do not contain structural information. Recently, besides the traditional flat data broadcast, a wealth of work

dealing with XML data broadcast has appeared. Some work addresses the performance optimization of query processing of XML streams in wireless broadcast [7, 18], while other work designs indexing techniques for XML data broadcast based on existing XML indexing techniques [19, 26]. However, their work mainly focuses on air indexing techniques similar to content based indexing techniques in XML stream processing or XPath query based indexing techniques. Moreover, this kind of work does not study the data placement problems in XML data broadcast.

The most related work is proposed in [20]. In that work, the broadcast schedules are generated based on clustering results of XML data on the server. However, the clustering process requires manually specifying the number of clusters and has to compare different clustering results based on clients' query distribution in order to find the optimal clustering result, which differs from our work in this paper.

7 Conclusion

In this paper, we have studied the data placement problem of periodic XML data broadcast. Taking advantage of the structured characteristics of XML data, we are able to generate effective broadcast programs based only on XML data on the server without any knowledge of the clients' access patterns. This not only makes our work distinguished from previous studies, but also enables it to have broader applicability. We presented a theoretical analysis of the problem and discussed structural sharing in XML data which forms the basis of our novel greedy data placement algorithm (GDPA). Our experiments demonstrated that the proposed algorithm could improve access efficiency and achieve better scalability.

In the future, we plan to further improve system's performance by investigating the insights of structural sharing among XML documents. For example, we may consider details on how to measure structural sharing distribution in an XML document set, how the distribution affects the expected access time of queries and how to choose a similarity measure based on structural sharing distribution in a set of XML documents.

References

1. Acharya, S., Alonso, R., Franklin, M.J., Zdonik, S.B.: Broadcast Disks: Data Management for Asymmetric Communications Environments. In: SIGMOD. (1995) 199–210
2. Acharya, S., Franklin, M.J., Zdonik, S.B.: Balancing Push and Pull for Data Broadcast. In: SIGMOD Conference. (1997) 183–194
3. Chang, Y.I., Hsieh, W.H.: An Efficient Scheduling Method for Query-Set-Based Broadcasting in Mobile Environments. In: ICDCS Workshops. (2004) 478–483
4. Chen, J., Lee, V.C.S., Liu, K.: On the Performance of Real-time Multi-item Request Scheduling in Data Broadcast Environments. Journal of Systems and Software **83**(8) (2010) 1337–1345

5. Chung, Y.D., Kim, M.H.: QEM: A Scheduling Method for Wireless Broadcast Data. In: DASFAA. (1999) 135–142
6. Chung, Y.D., Kim, M.H.: Effective Data Placement for Wireless Broadcast. Distributed and Parallel Databases **9**(2) (2001) 133–150
7. Chung, Y.D., Lee, J.Y.: An Indexing Method for Wireless Broadcast XML Data. Inf. Sci. **177**(9) (2007) 1931–1953
8. Diao, Y., Altinel, M., Franklin, M.J., Zhang, H., Fischer, P.M.: Path Sharing and Predicate Evaluation for High-Performance XML Filtering. ACM Trans. Database Syst. **28**(4) (2003) 467–516
9. Dice, L.R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3) (1945) 297–302
10. Ganesan, P., Garcia-Molina, H., Widom, J.: Exploiting Hierarchical Domain Structure to Compute Similarity. ACM Trans. Inf. Syst. **21**(1) (2003) 64–93
11. Helmer, S.: Measuring the Structural Similarity of Semistructured Documents Using Entropy. In: VLDB. (2007) 1022–1032
12. Imielinski, T., Viswanathan, S., Badrinath, B.R.: Data on Air: Organization and Access. IEEE Trans. Knowl. Data Eng. **9**(3) (1997) 353–372
13. IPTC: International Press Telecommunications Council, News Industry Text Format (NITF). http://www.nitf.org
14. Kang, S.H.: Wireless Data Broadcast Scheduling with Utility Metric Based on Soft Deadline. IEICE Transactions **94-B**(5) (2011) 1424–1431
15. Lee, G., Yeh, M.S., Lo, S.C., Chen, A.L.P.: A Strategy for Efficient Access of Multiple Data Items in Mobile Environments. In: MDM. (2002) 71–78
16. Lian, W., Cheung, D.W.L., Mamoulis, N., Yiu, S.M.: An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. IEEE Trans. Knowl. Data Eng. **16**(1) (2004) 82–96
17. Lin, D.: An Information-Theoretic Definition of Similarity. In: ICML. (1998) 296–304
18. Park, S.H., Choi, J.H., Lee, S.: An Effective, Efficient XML Data Broadcasting Method in a Mobile Wireless Network. In: DEXA. (2006) 358–367
19. Qin, Y., Sun, W., Zhang, Z., Yu, P., He, Z., Chen, W.: A Novel Air Index Scheme for Twig Queries in On-Demand XML Data Broadcast. In: DEXA. (2009) 412–426
20. Qin, Y., Wang, H., Sun, L.: Cluster-Based Scheduling Algorithm for Periodic XML Data Broadcast in Wireless Environments. In: AINA Workshops. (2011) 855–860
21. Qin, Y., Wang, H., Xiao, J.: Effective Scheduling Algorithm for On-Demand XML Data Broadcasts in Wireless Environments. In: ADC. (2011) 95–102
22. Rafiei, D., Moise, D.L., Sun, D.: Finding Syntactic Similarities Between XML Documents. In: DEXA Workshops. (2006) 512–516
23. Sun, B., Hurson, A.R., Hannan, J.: Energy-Efficient Scheduling Algorithms of Object Retrieval on Indexed Parallel Broadcast Channels. In: ICPP. (2004) 440–447
24. Sun, W., Liu, P., Wu, J., Qin, Y., Zheng, B.: An Automaton-Based Index Scheme for On-Demand XML Data Broadcast. In: DASFAA (2). (2012) 96–110
25. Sun, W., Shi, W., Shi, B., Yu, Y.: A Cost-Efficient Scheduling Algorithm of On-Demand Broadcasts. Wireless Networks **9**(3) (2003) 239–247
26. Sun, W., Yu, P., Qin, Y., Zhang, Z., Zheng, B.: Two-Tier Air Indexing for On-Demand XML Data Broadcast. In: ICDCS. (2009) 199–206
27. Sun, W., Zhang, Z., Yu, P., Qin, Y.: Efficient Data Scheduling for Multi-item Queries in On-Demand Broadcast. In: EUC (1). (2008) 499–505
28. Wu, J., Liu, P., Gan, L., Qin, Y., Sun, W.: Energy-Conserving Fragment Methods for Skewed XML Data Access in Push-Based Broadcast. In: WAIM. (2011) 590–601

29. Xu, J., Lee, D.L., Hu, Q., Lee, W.C. In: Handbook of Wireless Networks and Mobile Computing. John Wiley & Sons, Inc. (2002) 243–265

Appendix

A. Proof of Proposition 1

Proof. According to the two constraints of function $f(\mathbf{X})$, we have

$$\begin{aligned} f(\mathbf{X}) &= (x_1 + x_2 + \ldots + x_n)^2 - 2 \cdot \sum_{i \neq j} x_i \cdot x_j \\ &\leq M^2 \end{aligned}$$

When $x_1 = x_2 = \ldots = x_{k-1} = 0$ and $x_k = M$ (or any other kind of combinations like this, which means we have only 1 positive variable[2].), $f(\mathbf{X})$ reaches its upper bound $\overline{f(\mathbf{X})} = M^2$. In all other cases, if two or more variables are positive, i.e. $x_1 > 0$ and $x_2 > 0$, we have $2 \cdot \sum_{i \neq j} x_i \cdot x_j \geq 2 \cdot (x_1 \cdot x_2) > 0$ which indicates $f(\mathbf{X}) < M^2$.

B. Proof of Proposition 2

Proof. We utilize mathematical induction to prove it (for all $k \geq 1$).

For the base step, when $k = 1$, $f(\mathbf{X}) = x_1^2$, it is trivial to prove that $f(\mathbf{X}) = M^2 \geq 1 \cdot (\frac{M}{1})^2$ and $\underline{f(\mathbf{X})} = M^2$ since $x_1 = M = \frac{M}{1}$.

For the inductive step, we assume that when $k = n$, $f(\mathbf{X}) \geq n \cdot (\frac{M}{n})^2$ and when $x_1 = x_2 = \ldots = x_n = \frac{M}{n}$, $f(\mathbf{X})$ reaches its lower bound $\underline{f(\mathbf{X})} = n \cdot (\frac{M}{n})^2$. Then for $k = n + 1$, we have

$$\begin{aligned} f(\mathbf{X}) &= \sum_{i=1}^{n} x_i^2 + x_{n+1}^2 \\ &\geq n \cdot (\frac{M - x_{n+1}}{n})^2 + x_{n+1}^2 \\ &= \frac{(n+1) \cdot (x_{n+1} - \frac{M}{n+1})^2 + \frac{n \cdot M^2}{n+1}}{n} \\ &\geq (n+1) \cdot (\frac{M}{n+1})^2 \end{aligned}$$

From the above induction, we can see that when $x_{n+1} = \frac{M}{n+1}$ and $x_1 = x_2 = \ldots = x_n = \frac{M - x_{n+1}}{n} = \frac{M}{n+1}$, $f(\mathbf{X})$ reaches its lower bound $\underline{f(\mathbf{X})} = (n+1) \cdot (\frac{M}{n+1})^2$.

Because we have shown both the base step and the inductive step, by the principle of mathematical induction the proposition is true.

[2] Note that we cannot have all variables to be 0 since M is positive.

C. Proof of Proposition 3

Proof. According to our definitions of $f(\mathbf{X})_{x_{m+1}}$, we have

$$f(\mathbf{X})_{x_{m+1}} = \sum_{i=1}^{m} x_i^2 + x_{m+1}^2 + \sum_{i=m+2}^{k} x_i^2$$

Since $M' = M - \sum_{i=1}^{m} x_i = \sum_{i=m+1}^{k} x_i$, according to Proposition 1, we have

$$\begin{aligned} \underline{f(\mathbf{X})}_{x_{m+1}} &= \sum_{i=1}^{m} x_i^2 + x_{m+1}^2 + (M' - x_{m+1})^2 \\ &= \sum_{i=1}^{m} x_i^2 + 2 \cdot (x_{m+1} - \frac{M'}{2})^2 + \frac{M'^2}{2} \end{aligned}$$

According to the above result, for any $x_{m+1} < x'_{m+1}$ we can infer that

1. $\overline{f(\mathbf{X})}_{x_{m+1}} > \overline{f(\mathbf{X})}_{x'_{m+1}}$, if $x_{m+1} < x'_{m+1} \leq \frac{M'}{2}$
2. $\overline{f(\mathbf{X})}_{x_{m+1}} < \overline{f(\mathbf{X})}_{x'_{m+1}}$, if $\frac{M'}{2} \leq x_{m+1} < x'_{m+1}$
3. Indefinite for all other cases

D. Proof of Proposition 4

Proof. According to our definitions of $f(\mathbf{X})_{x_{m+1}}$, we have

$$f(\mathbf{X})_{x_{m+1}} = \sum_{i=1}^{m} x_i^2 + x_{m+1}^2 + \sum_{i=m+2}^{k} x_i^2$$

Since $M' = M - \sum_{i=1}^{m} x_i = \sum_{i=m+1}^{k} x_i$, according to Proposition 2, we have

$$\begin{aligned} \underline{f(\mathbf{X})}_{x_{m+1}} &= \sum_{i=1}^{m} x_i^2 + x_{m+1}^2 + \\ &\quad (k-m-1) \cdot (\frac{M' - x_{m+1}}{k-m-1})^2 \\ &= \sum_{i=1}^{m} x_i^2 + \\ &\quad \frac{(k-m)(x_{m+1} - \frac{M'}{k-m})^2 + \frac{(k-m-1)M'^2}{k-m}}{k-m-1} \end{aligned}$$

According to the above result, for any $x_{m+1} < x'_{m+1}$ we can infer that

1. $\underline{f(\mathbf{X})}_{x_{m+1}} > \underline{f(\mathbf{X})}_{x'_{m+1}}$, if $x_{m+1} < x'_{m+1} \leq \frac{M'}{k-m}$
2. $\underline{f(\mathbf{X})}_{x_{m+1}} < \underline{f(\mathbf{X})}_{x'_{m+1}}$, if $\frac{M'}{k-m} \leq x_{m+1} < x'_{m+1}$
3. Indefinite for all other cases