Semantic XML Filtering on Peer-to-Peer Networks using Distributed Bloom Filters

Panagiotis Antonellis, Stavros Kontopoulos, Christos Makris, Yannis Plegas and Nikos Tsirakis
Department of Computer Engineering and Informatics, University of Patras, Patras, Greece

Keywords: XML Filtering, Distributed Processing, Peer-to-Peer Networks, Bloom Filters, Semantic Filtering, Word Sense Disambiguation.

Abstract: Information filtering systems constitute a critical component in modern information seeking applications. As the number of users grows and the information available becomes even bigger it is imperative to employ scalable and efficient representation and filtering techniques. Typically the use of XML representation entails the profile representation with the use of the XPath query language and the employment of efficient heuristic techniques for constraining the complexity of the filtering mechanism. However, as the number of XML documents exchanged daily grows rapidly, the need for distributed management is becoming vital. In this paper we introduce the Distributed Bloom Filters and we propose a new distributed XML filtering system for peer-to-peer (P2P) networks. The major advantage of Distributed Bloom Filters, in comparison to the classical structure is their space efficiency and improved performance. The proposed system efficiently filters the incoming XML documents using a virtual index created on top of the network. In addition, the proposed system supports semantic disambiguation of both the stored user profiles and the XML documents, thus providing better matching results.

1 INTRODUCTION & RELATED WORK

Information filtering systems (Aguilera et al., 1999) are systems that provide two main services: document selection and document delivery. Lately, there have appeared (Antonellis et al., 2009), (Miliaraki and Koubarakis, 2010), (Ning and Liu, 2010) a number of systems that use XML representations for both documents and user profiles and that employ various filtering techniques to match the XML representations of user documents with the provided profiles. Among the growing amount of objects shared by P2P applications there is an increasing number of XML-documents that is being shared among peers. There are a number of search engines for P2P networks such as the DHT-based systems of (Bender et al., 2005) and (Podnar et al., 2007). However, none of these approaches supports XML-Retrieval techniques.

In this work, we introduce the idea of Distributed Bloom Filters, which utilize the fast lookups of Bloom Filters (Bonomi et al., 2006a), (Bonomi et al., 2006b), with the advantage of the distributed storage which reduces the storage overhead in each network peer and at the same time improves the performance. Based on the Distributed Bloom Filters, we present a new P2P system that supports semantic filtering of the incoming XML documents. The main contributions of our work are:

- Introduction of Distributed Bloom Filters and efficient indexing using them.
- Efficient distribution of the user profiles in the network's peers.
- Word disambiguation of the tags of the stored user profiles and the incoming XML documents for supporting semantic filtering.

2 DISTRIBUTED BLOOM FILTER

Let *BF* be a Bloom Filter of *m* bits. In order to reduce the space overhead per peer and also improve the membership check performance, we introduce the idea of Distributed Bloom Filter. The bloom filter will be stored in *p* peers, so we cut the bit array of the bloom filter in *p* segments with *c* bits each.

Antonellis P., Kontopoulos S., Makris C., Plegas Y. and Tsirakis N..
Semantic XML Filtering on Peer-to-Peer Networks using Distributed Bloom Filters.
DOI: 10.5220/0004363301330136
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 133-136
ISBN: 978-989-8565-54-9

Let $O=\{o_1,o_2,..,o_n\}$, $|O|=n$, be the set of the objects that are tested for membership by the bloom filter data structure. Let $h_1,h_2,...,h_k$ be the set of hashing functions used by the bloom filter. For each object *j* we define the ordered sequence H_j as: $H_j(O_j)=\{h_1(O_j),h_2(O_j),...,h_k(O_j)\}$.

We now split the result bit sequence of each hash function h_i into *p* segments and identify which segments contain at least one bit set to 1. Based on this, we define the following sequences:

$$hs_i(O_j)=\underbrace{(b_1^i b_2^i \ldots b_p^i)}_{p\ bits}$$

, where b_ℓ^i is set to 1 if the $h_i(O_j)$ sequence sets at least one bit of the ℓ *-th* segment, otherwise is set to 0. From that set we can define the ordered sequence:

$$W_j=\{hs_1(O_j),hs_2(O_j),\cdots,hs_k(O_j)\}$$

Next we apply a hashing function on W_j:

$$hash(W_j=\{hs_1(O_j),hs_2(O_j),\cdots,hs_k(O_j)\})=IDj\ .$$

We use the *IDj* as an indexing value for an object on a DHT based network. After inserting all the objects into the Distributed Bloom Filter, the bit sequence is split into *p* segments and distributed into *p* different peers to reduce the space overhead and also increase the speed efficiency. We call this new data structure: *M-DBF*.

3 SEMANTIC XML FILTERING ON P2P NETWORKS

We employ the Distributed Bloom Filter in order to design a new distributed semantic XML Filtering system, which can work on top of any DHT peer-to-peer network. Moreover we enhance the whole scheme by embedding semantic techniques based on WordNet (Miller et al. 1990). The proposed system works as follows:

- It clusters the user profiles using the k-Means algorithm.
- It distributes the user profiles in the network's peers based on the belonging cluster.
- It utilizes a multi-level index, following the approach described in (Antonellis *et al*, 2009), based on M-DBFs
- It performs word sense disambiguation of the tags of the stored user profiles and incoming XML documents.

3.1 Profile Clustering and Distribution

The clustering of the initial set of user profiles is performed once in a central server, before initializing our system. The utilized clustering algorithm is the k-Means algorithm in conjunction with the distance metric described in (Antonellis and Makris, 2008b) for calculating the distance between a pair of user profiles.

The underlying P2P network is divided into neighbourhoods, with each neighbourhood storing the user profiles of a single cluster, and with each neighbourhood consisting of physical neighbour peers. In order to optimize the filtering performance as well as the update operations on the stored user profiles, each neighbourhood is organized in a two-level hierarchy, as described in the Section 3.2. For every formed cluster, an M-DBF is constructed, called *Cluster MDBF*, that stores all the distinct paths contained in the user profiles of that cluster.

3.2 Distributed Indexing Scheme

The proposed system utilizes a distributed hierarchical indexing system, as seen in figure 1. Each network neighbourhood is consisted of physically close nodes and it is responsible for storing and handling the user profiles of a single cluster.

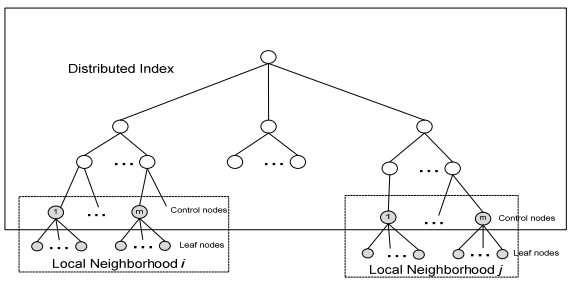

Figure 1: Distributed Indexing Scheme.

The low-level peers of a neighbourhood N_i are called *leaf peers* and are used for actually storing the XML documents of the *i*-th cluster, while the top level peers are called *control peers* and are used for query routing through the current neighbourhood as well as through different neighbourhoods in the network.

A control peer is responsible for a subset of the leaf peers in N_i, called its leaf *subset peers* and the total number of control peers is much smaller than the total number of leaf peers. The control peers of each neighbourhood know all their leaf subset peers and can redirect any query to all of them. On the other hand, the leaf peers know only their control peer as well all their sibling leaf peers. All the control peers of the network are organized in a multi-level indexing scheme inspired by VBI-tree (Jagadish et al., 2006).

3.3 XML Document Filtering

The XML filtering process of the proposed system utilizes the previously described indexing structure to efficiently forward the incoming XML documents to the appropriate control peers of the neighbourhoods that are possible to match the document. When an XML document is submitted to a peer p_j of the network, the peer p_j is automatically responsible for processing and filtering the submitted document. Firstly, it checks its LBF to see if the XML document is likely to match with any of its user profiles. If so, it performs a full filtering against all its stored user profiles using the XFIS (Antonellis and Makris, 2008a) filtering algorithm and stores the results in its cache. Then, it forwards the XML document to its parent control peer for further routing. The control peer uses the VBI-tree to forward the XML document to any other control peer which its M-DBF matches with the XML document.

Every peer that matches any of its user profiles with the incoming XML document propagates the results back to the original peer, because this peer is responsible for gathering the total filtering results.

3.4 Word Sense Disambiguation

In the proposed system, the textual information of XML documents is semantically-enriched with the support of WordNet (Miller *et al*, 1990). WordNet is a lexical online ontology including over 110,000 concepts. The related concepts are grouped into sets of synonyms which are called *synsets*. Each synset represents a lexical concept and is described by a short textual description the gloss. A synset represents a lexical meaning, or sense, which can be assigned to multiple terms.

Our main purpose is the *word sense disambiguation* of the tags occurring in a given path, and the assignment of unique senses to each tag. That way, our system will be able to find related user profiles with the incoming document, even though they do not use exactly the same tags. Word sense disambiguation (WSD) governs the process of identifying which sense of a word is used, when the word has multiple meanings. Initially, we identify all the different senses of a term. The next step is the selection of the most appropriate sense for the respective term. The process which was followed was a dictionary-based WSD and was handled using the lexical knowledge base of WordNet. WordNet provides the texts of the senses definitions (glosses) and gives us the opportunity to adapt the assumption that the most plausible sense to assign to a term with multiple senses is the one that maximizes the semantic relatedness among the senses.

The semantic similarity between two senses is computed using two different approaches. The first approach employs as similarity metric the Wu and Palmer (Budanitsky and Hirst, 2006):

$$similarity(c_i, c_j) = \frac{2depth(LCA(c_i, c_j))}{depth(c_i) + depth(c_j)}$$

The second approach which is also used in (Tagarelli et al., 2009) has been formalized in a measure of semantic relatedness between word senses based on the notion of *extended gloss overlap* (Patwardhan et al., 2003), and has the merit of considering phrasal matches and weighting them more heavily than single word matches.

4 EXPERIMENTS

We have built a prototype P2P emulator to evaluate the performance of our proposed filtering system over large-scale networks.

4.1 Varying Number of Network Peers

In this experiment, we wanted to study the relationship between the number of peers in the network and the number of hops required for each query to be processed. Thus, we created 8 clusters of totally 1500000 user profiles which were distributed in 50000, 100000, 200000 and 500000 peers in the network. For each case we counted the average number of hops for each query in the query set. The experimental results are shown in Table 1.

4.2 Semantic Disambiguation of XML Tags

In this experiment, we evaluated the precision and

speed of the proposed semantic disambiguation during XML filtering, in comparison with the technique described in (Tagarelli et al., 2009). We have used a set of 50000 and 100000 stored user profiles and 45000 incoming XML documents for testing. The results are displayed in Table 2.

Table 1: Number of hops.

#Peers	#Total hops	#Neighbour hops	#Routing hops	Perc.
50000	6834	5391	1443	13.6%
100000	10351	8280	2071	10.3%
200000	12179	9778	2401	6.1%
500000	14260	11323	2937	2.8%

Table 2: Precision and time of XML tag disambiguation (A: our approach, B: Tagarelli's approach).

#Profiles	Precision A	Time A	Precision B	Time B
5000	85%	1291s	90%	1972s
10000	82%	2189s	88%	3281s

As we can easily observe, our approach performs very well, achieving an average precision of about 84%. In addition, due to the simplicity of our approach, the required time for the disambiguation is much smaller (about 30% faster).

AKNOWLEDGEMENTS

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

This research has been co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)-Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

REFERENCES

Abiteboul, S., Manolescu, I., Polyzotis, N., Preda, N. and Sun, C. XML processing in DHT networks. *ICDE*, 2008.

Antonellis, P. and Makris, C. XFIS: An XML filtering system based on string representation and matching. *International Journal on Web Engineering and Technology (IJWET)*, 4(1), 70-94, 2008.

Antonellis, P. and Makris, C. XML Filtering Using Dynamic Hierarchical Clustering of User Profiles. *DEXA*, 537-551, 2008.

Antonellis, P., Makris, C. and Tsirakis, N. Utilizing XML Clustering for Efficient XML Data Management on P2P Networks. *DEXA*, 68-82, 2009.

Aguilera, M. K., Strom, R.E., Stunnan, D. C., Astley, M. and Chandra, T. D. Matching events in a content-based subscription system. *PODC*, 53–61, 1999.

Bender, M., Michel, S., Weikum, G. and Zimmer, C. The MINERVA Project - Database Selection in the Context of P2P Search. *BTW Conference*, 2005.

Bonomi, F., Mitzenmacher, M., Panigrahy R., Singh S. and Varghese G. An Improved Construction for Counting Bloom Filters. *ESA*, 684-695, 2006.

Bonomi, F., Mitzenmacher M., Panigrahy R., Singh S. and Varghese G. Beyond bloom filters: from approximate membership checks to approximate state machines. *SIGCOMM* , 315-326, 2006.

Budanitsky, A. and Hirst, G. Evaluating WordNet-based measures of lexical semantic relatedness. *Association for Computational Linguistics*, 32, 32-47, 2006.

Jagadish, H. V, Ooi, B. C., Vu, Q. H, Zhang, R. and Zhou. A. VBI-Tree: a Peer-to-Peer Framework for Supporting Multi-Dimensional Indexing Schemes. *ICDE*, 2006.

Miliaraki, I. and Koubarakis, M. Distributed structural and value XML filtering. *4th ACM International Conference on Distributed Event-Based Systems*, 2–13, 2010.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D. and Miller. K. WordNet: An online lexical database. *Int. J. Lexicograph*, 3(4), 235-244, 1990.

Ning, B. and Liu, C. XM filtering with XPath expressions containing parent and ancestor axes. *Information Sciences*, Elsevier, 210 (Nov. 2010), 41-54, 2010.

Patwardhan, S., Banerjee, S., and Pedersen, T. Using measures of semantic relatedness for word sense disambiguation. *In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 241–257, 2003.

Podnar, I., Rajman, M., Luu, T., Klemm, F. and Aberer, K. Scalable Peer-to-Peer Web Retrieval with Highly Discriminative Keys. *ICDE*, 2007.

Tagarelli, A. and Greco, S. Semantic clustering of xml documents. *ACM Transactions on Information Systems,* 28 (1), 1-56, 2010.

Tagarelli, A., Longo, M. and Greco S. Word Sense Disambiguation for XML Structure Feature Generation. *In Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, 2009.