18th International Workshop on Database and Expert Systems Applications

Dimensionality Reduction in a P2P System

Mouna Kacimi
Max-Planck Institut für Informatik
66123 Saarbrücken
Germany

Kokou Yétongnon
Laboratoire LE2I
University of Bourgogne
Dijon, 21078 Cedex France

***Abstract*— Peers and data objects in the Hybrid Overlay Network (HON) are organized in a n-dimensional feature space. As the dimensionality increases, peers and data objects become sparse and the distance measures become increasingly meaningless which leads to serious problems affecting HON performance. In this paper we propose a distributed feature selection technique reduce the dimensionaliy in HON. We study in our simulations the impact of the proposed feature selection technique on query results quality and show that it achieves high recall and precision.**

I. Introduction

In recent years, content-based retrieval of high dimensional data has been a challenging problem in several fields such as data analysis and data mining. A number of applications including multimedia and text retrieval require the use of high dimensional methods to provide capabilities for finding data objects which are similar by content. In our work, we focus on content-based and similarity search in P2P networks. We have proposed a Hybrid Overlay Network (HON) [12] that organizes both peers and data in an n-dimensional feature space based on content description. The basic idea is to define a partition of the feature space into cells and use the distribution of data objects over the cells as the basis for defining peer similarity, creating clusters and computing query similarities to peers and clusters. As the dimensionality of the feature space increases, the data objects are sparsely distributed over the space resulting in several problems. First, the number of cells increases exponentially with increasing dimensionality. For example, using a 30 dimensional data and only 2 partitions for each feature, we can have more than one billion of cells. Thus, the computation time and the storage space needed to store cells description increases dramatically. Second, a cluster might be divided into a large number of cells and many or even all these cells might have a density less than the required threshold. In addition, the clusters might only exist in subsets of high dimensional spaces. Since the number of possible subspaces is also exponential in the dimensionality of the space, clusters cannot be easily defined.

Several dimensionality reduction techniques have been proposed to address the curse of dimensionality [2], [3], [8], [11], [14], [14]. These techniques aim to project data objects from a high dimensional space to a lower dimensional space. The resulting subspace is described by a set of new features that are the combination of the original features. Most of these techniques are centralized. They are called GDR techniques (Global Dimensionality Reduction) because they use the whole dataset for the projection. In a P2P context, the dataset is highly distributed among peers. Therefore, data need to be collected in a central server to do the projection which might limit the use of GDR techniques in P2P systems.

It has been shown in [6] that GDR techniques provide good results only if the dataset is globally correlated which means that the variation in the data can be captured by few dimensions. In practice, datasets are often not globally correlated. Thus, GDR techniques lead to a significant loss of information. To address this problem, Local Dimensionality Reduction (LDR) techniques [6] have been applied individually to clusters of locally correlated data which results in a different subspace for each cluster. In P2P systems, peers belonging to different subspaces need to communicate. Hence, mapping techniques have to be introduced. Since the features that describe the new subspaces are meaningless to the user because they are a combination of the original features, the mapping might require a complex schema.

Dimensionality reduction techniques presented above can be hardly used in P2P systems, mainly because of their centralized aspect. Thus, some efforts have been made to propose distributed dimensionality reduction techniques [1], [15] where each node in the network sends a sample of data representing its content to a central server. This server projects the set of samples in a new subspace and sends its description to all nodes of the network. When the nodes receive that description, they project all their data in the new subspace. Actually, these techniques have a distributed input but a centralized processing. In addition, the resulting subspace might provide a loss of information since data are usually not globally

1529-4188/07 $25.00 © 2007 IEEE
DOI 10.1109/DEXA.2007.85

correlated.

The focus of this paper is to address the dimensionality issue by viewing a P2P systems as one or more overlay networks. Each overlay is described by a specific and limited number of features. To guide the selection of features for the overlays, we propose a distributed feature selection technique that describes dense regions of the feature space. Its is a variant of the weighted feature selection algorithm proposed by Wang et al [16]. Each peer selects the set of features describing most of its data objects and joins one or multiple overlay networks depending on its selected features. The main contributions of this work are threefolds:

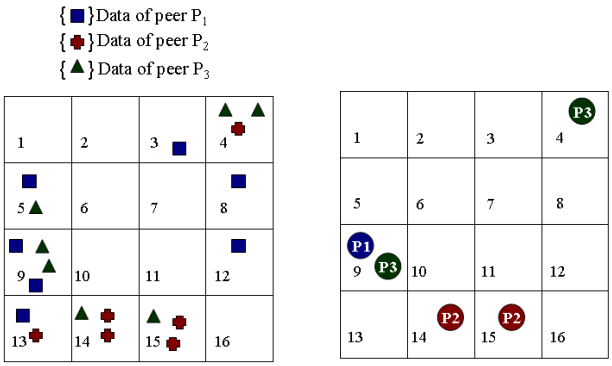
1) we propose a weighted feature selection technique to reduce the dimensionality of HON improving its efficiency.
2) we evaluate the performance of the feature selection technique using different data and query distributions, namely, uniform and Zipfian.
3) we improve the performance of the proposed feature selection technique by considering discarded features to increase the recall and the precision.

The remainder of the paper is organized as follows. In the next section, we give an overview of dimensionality reduction techniques. In section 3, we give a brief description of HON. Section 4 presents the weighted feature selection technique. Section 5 presents the evaluation results. And finally section 6 concludes the paper.

II. Related Work

Dimensionality reduction techniques can be classified into two categories: *Feature Construction* and *Feature Selection*. Feature construction approaches project points from a higher dimensional space to a subspace having a lower dimensionality. By contrast, feature selection techniques consider that not all features are important in describing data objects. Therefore, they select a subset of the original features that describes most of the data.

There are several feature construction techniques that have been proposed for different problems. Carrerira-Peripinan classifies in [5] dimensionality reduction problems into three categories: *Hard*, *Soft* and *Visualisation*. In *Hard* dimensionality reduction problems, the dimension of data objects ranges from hundreds to hundreds of thousands of features. This category of problems includes pattern recognition and classification studies involving images or audio data types. In practical cases, the most widespread technique used for *Hard* problems is Principal Components Analysis (PCA). The *Soft* category of dimensionality problem includes data objects having



(a) Data organization (b) Peers organization

1: Hybrid Overlay Network

less than tens of features. Most statistical analysis in fields such as social science and psychology fall into this category. The number of these features is never too-high which makes the dimensionality reduction not very drastic. Soft problems also employ the PCA algorithm and other techniques such as Factor Analysis [8], Discriminant Analysis [2] and Multidimensional Scaling [14]. The last category of reduction problems is the *visualisation* where the data objects are not high dimensional, but they need to be projected in 2, 3 or 4 dimensional spaces in order to plot them. For *Visualisation* purpose many methods have been used in practice including PCA, Projection Pursuit [11], Multidimensional Scaling [14], and Self-Organizing maps [3] including their variants.

Feature selection has been the focus of interests of many fields and applications such as data mining [7], machine learning [4], pattern recognition [9], text categorization [8] and image retrieval [17]. The main idea of feature selection is to reduce the number of features by removing irrelevant, redundant, or noisy information to improve the application performance such as computational time and result comprehensibility.

III. Hybrid Overlay Network

The Hybrid Overlay Network (HON) organizes peers and data to perform an efficient similarity search based on range and nearest neighbor queries. The data contents of peers are represented by n-element feature vector, where each element is a particular feature or attribute associated with data object (e.g., color for an image, concept or key word for a text document). Since each data object is described by a feature vector, it can be seen as a

point in a n-dimensional feature space.

The feature space is described by n features $f_1, f_2, ...f_n$. The basic idea is to define a partition of the feature space into cells and use the distribution of data objects over the cells as the basis for defining peer similarity, creating clusters and computing query similarities to peers and clusters. Thus, two peers are considered similar if their contents are distributed on the same sub regions of the feature space. Figure 1a shows the partition cells of a 2-dimensional feature space using the features f_1 and f_2. Data of peers P_1, P_2 and P_3 are distributed over cells according to their description. The number of data objects in each cell is recorded using a **cell density** measure. This notion of cell density is used for mapping a peer to a cell, to create peer clusters and to reduce the dimensionality of the feature space. Peers are mapped into cells according to the distribution of their data objects. using a threshold value, a peer is mapped to a cell if the number of its data objects in that cell is higher than T. Figure 1b shows an example of peer mapping where $T = 1$.

Once data and peers are mapped into the feature space, clusters of peers are created according to cells' density to provide a high recall. Grouping peers within cells having high densities increases the number of retrieved similar objects. Clustering will not be discussed further in this paper. More details about the density-based algorithm and query processing can be found in our previous work [13].

IV. Dimensionality Reduction in HON

To reduce the dimensionality in HON, we use a variant of the *Weighted Feature Selection* algorithm proposed by Wang et al. [16] which aims to (1) cluster data point using k-means algorithm [10] and (2) extract the relevant features for each cluster using histogram analysis to assign greater weight to relevant features. In our work, the selection of features is processed in each peer without any clustering consideration. Each peer runs locally the *Weighted Feature Selection* technique to select the most relevant features to its data objects. Depending on their selected features, peers joins one or multiple overlay networks, where each overlay network is associated to a subset of features that describe common set of peers.

Let $f_1, ...f_n$ be the set of features describing the feature space. We have defined a partition of the feature space into cells by dividing the range of values $[f_i^{min}, f_i^{max}]$ of a feature f_i into m_i intervals of size $\lceil \frac{f_i^{max} - f_i^{min}}{m_i} \rceil$, for $i = 1, 2...n$. We assume that the denser is the distribution of a given feature f_i, the greater is the probability that f_i is the dominant feature in representing the dataset. Considering this assumption, each peer starts by computing for each feature f_i the number of its data objects in each partition. Let D_{ij} be the number of the peer's data objects in the partition P_{ij} of the feature f_i. We define a region of a feature f_i by:

$$FeatureRegion_{(i)} = \sum_{j=1}^{m_i} D_{ij} \times \frac{f_i^{max} - f_i^{min}}{m_i}$$

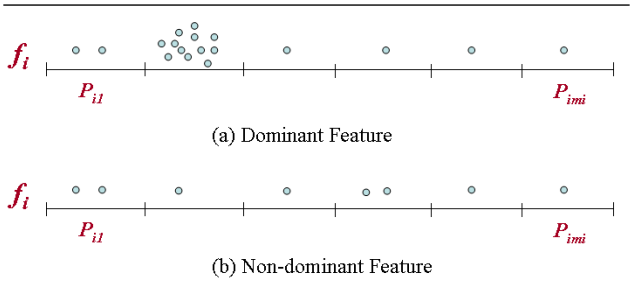
The density for the feature f_i is defined by:

$$FeatureDensity_{(i)} = 1 - \frac{FeatureRegion_i}{Max(D_{ij}) \times (f_i^{max} - f_i^{min})}$$

The $FeatureDensity_{(i)}$ represents the density value of the distribution of the feature f_i. Following a uniform distribution, the number of data objects in each partition P_{ij} is almost the same as shown in figure 2b. Consequently the ratio $\frac{FeatureResgion_i}{Max(D_{ij}) \times (f_i^{max} - f_i^{min})}$ tends to 1 resulting in a low feature density. By contrast, if the data objects are distributed using a Zipfian low where they are highly concentrated in few partitions as shown in figure 2a, the feature density value becomes higher. Therefore, the larger is $FeatureDensity_{(i)}$, the denser is the value distribution for f_i. The $FeatureDensity_{(i)}$ values are then used to compute the weight that we associate to each feature. This weight value will indicate the relevancy of features. Let $\{w_i, ..., w_n\}$ be the corresponding weights to the features $\{f_1, ...f_n\}$. We define the weight w_i of the feature f_i as:

$$w_i = \frac{FeatureDensity_{(i)}}{\sum_{j=1}^{n} FeatureDensity_{(j)}}$$

Two different strategies can be used by a peer to select the important features in describing its data objects. The first strategy uses a threshold value TW. If the weight w_i of a feature f_i is higher than the threshold value TW, then the feature f_i is be selected. Otherwise, it is be discarded. The second strategy consists in defining the number k of features that have to be selected. In this case, the peer ranks



2: Feature density

the features according to their weight values. Then, only the k first features will be selected as the most important ones.

V. Evaluation

We have evaluated the weighted feature selection technique that we proposed to reduce dimensionality in HON. An efficient dimensionality reduction technique should preserve data information as much as possible. Concretely, we use three metrics for the evaluation: *Recall*, *Precision* and *F-Measure* described as follows:

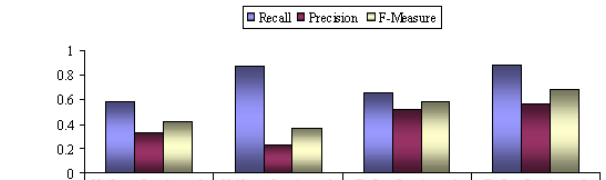
1) extitRecall r: represents the percentage of retrieved responses out of the available responses in the network.
2) *Precision p*: represents the percentage of relevant responses to the query out of the retrieved responses.
3) *F-Measure*: is a weighted harmonic mean of precision and recall. In our simulation we use an F_1 measure that gives the same weight to the recall r and the precision p. F_1 is given by:

$$F_1 = \frac{2rp}{r+p}$$

We consider in our simulation a global set of features that describe peers' data objects. Each peer applies locally the weighted feature selection algorithm to select the important features in describing its content. The number of selected features varies from a peer to another depending on the distribution of its data objects. Two types of data distribution are used in this simulation: uniform data distribution and Zipfian data distribution. Recall that using a uniform distribution, peers' data objects have equal chance to be mapped to any cell of the feature space. By contrast, using a Zipfian distribution, data objects are mapped to few cells of the feature space. Note that data objects of each peer follow a different Zipfian distribution.

We simulate 1000 peers described by 30 common features. When peers select their relevant features using the weighted feature selection algorithm, they initiate $500,000$ queries to evaluate the quality of the search in the reduced spaces by computing the average recall, precision and F-measure. To study the behavior of the feature selection algorithm according to data distributions, we run four different simulations corresponding to four possible cases presented in the following:

In the first simulation, queries follow a uniform distribution where peers send queries randomly to cells. Figure 3 shows that using a uniform distribution for peers' content, the average recall is equal to 58% and the precision is equal to 32%. This means



	Uniform Queries and Uniform Data	Uniform Queries and Zipfian Data	Zipfian Queries and Uniform Data	Zipfian Queries and Zipfian Data
Recall	0.586745632	0.878945124	0.659821212	0.886543215
Precision	0.328329914	0.230077638	0.518609519	0.561307956
F-Measure	0.421049702	0.364691555	0.580754647	0.687396288

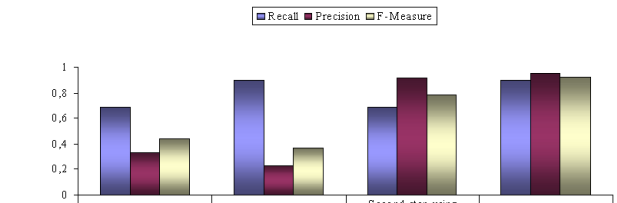
3: Weighted feature selection performance

that 58% of the relevant answers for a given query are retrieved using the reduced spaces. In addition, the relevant answers correspond to 32% of the total answers. Thus, 68% of unnecessary messages related to irrelevant answers are generated over the network. The F-measure in this case is equal to 42% indicating the non efficiency of the search. Note that in the second simulation, a Zipfian distribution increases the recall to 87% but do not improve the performance providing 23% of precision as shown in figure 3.

In the third and the fourth simulations, queries follow a Zipfian distribution where peers send queries to few cells in the feature space. Figure 3 shows that the average precision and F-measure is significantly improved using a Zipfian distribution of queries. For example, using a uniform distribution of peers' content, the average precision is equal to 51%. Thus, a reduced amount equal to 49% of unnecessary messages is generated over the network. The F-measure in this case is equal to 58% indicating a better efficiency than the first two simulations. By contrast, using a Zipfian distribution of peers' content provides the highest performance of 88% of recall, 56% of precision and 68% of F-measure.

According to the results presented above, a Zipfian distribution of data objects and queries performs the weighted feature selection algorithm comparing to uniform distributions. This can be explained by the fact that data objects of peers following a Zipfian distribution fall into few cells in the feature space. Therefore, the weighted feature selection algorithm can efficiently determine the most relevant features to peers content comparing to uniform distribution. A uniform distribution maps peers' data objects randomly to cells. Subsequently, a larger amount of information is lost resulting in a low precision and recall.

To provide more precise search, we propose that each peer makes use of its discarded features. The relevant features are used for assigning peers to over-



	Firststep using Relevant Features (Uniform Distribution)	First Step using Relevant Features (Zipfian Data Distribution)	Second step using Discarded Features (Uniform Data Distribution)	Second step using Discarded Features (Zipfian Data Distribution)
Recall	0,687921296	0,898977349	0,687921296	0,898977349
Precision	0,328329914	0,230077638	0,917234523	0,957234523
F-Measure	0,444506511	0,366385318	0,786198018	0,927191736

4: Precise search using discarded features

lays, organizing peers and their data in the feature space, creating clusters and routing queries. By contrast, discarded features can be used when computing similarities between queries and data objects to provide a precise search. Consider a query Q described by the set of features F_Q. When a peer receives the query Q, it computes the similarity between Q and each of its data objects in two steps:

1) In the first step, the peer computes the similarity using its relevant features that are common with the query Q and builds a list S of the similar objects to the query Q.
2) In the second step, the peer computes the similarity between each object of the list S and the query Q using the discarded features. The peer considers the discarded features that describe the query Q. The goal is to remove false positives and to keep as much as possible the relevant answers to the query Q

We have run a set of experiments using the same previous configuration to evaluate the efficiency of the search when taking advantage of discarded features. Figure 4 measures the search performance at the end of each step. When using a uniform distribution, the search using relevant features provides 32% of precision. At the end of the second step that selects the relevant answers using the discarded features, the precision increases to 91% performing the search efficiency. We can notice that the Zipfian distribution of data objects provides the highest precision of 95%.

VI. Conclusion

We have discussed high dimensionality problems related to the Hybrid Overlay Network that organizes data and peers into similar clusters. We have proposed a variation of a weighted feature selection based on a filter technique. It uses a goodness criterion depending on cells densities and threshold values to reduce the dimensionality by eliminating insignificant features. Therefore, each peer reduces its dimensionality by selecting a subset of dominant features. The simulation results showed that the proposed feature selection technique provides a high recall and precision.

References

[1] F. N. Abu-Khzam, N. F. Samatova, G. Ostrouchov, M. A. Langston, and G. A. Geist, "Distributed dimension reduction algorithms for widely dispersed data," *Parallel and Distributed Computing and Systems PDCS*, 2002.
[2] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis - a brief tutorial," Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University, Tech. Rep., 1999.
[3] G. Barreto and A. A. ujo, "Time in self-organizing maps: An overview of models," *International Journal of Computer Research*, vol. 10(2), pp. 139–179, 2001.
[4] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.
[5] M. A. Carrerira-Peripinan, "A review of dimension reduction techniques," Departement of Computer Science - University of Sheffield, Tech. Rep. CS-96-09, 1997.
[6] K. Chakrabarti and S. Mehrotra, "Local dimensionality reduction: A new approach to indexing high dimensional spaces," *In Proceedings of the 26th VLDB Conference*, pp. 89–100, 2000.
[7] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering - a filter solution," *In Proceedings of the IEEE International Conference on Data Mining (ICDM02)*, pp. 115–124, 2002.
[8] J. DeCoster, "Overview of factor analysis." Department of Psychology - University of Alabama, 348 Gordon Palmer Hall - Box 870348, Tech. Rep., 1998.
[9] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, March 1997.
[10] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," *Prentice Hall*, 1988.
[11] L. O. Jimenez and D. Landgrebe, "High dimensional feature reduction via projection pursuit," School Of Electrical and Computer - Prudue University, West Lafayette In 47907-1285, Tech. Rep. TR-ECE 96-5, 1995.
[12] M. Kacimi, Y. Ma, R. Chbeir, and K. Yetongnon, "Hon-p2p: A cluster-based hybrid overlay network for multimedia object management," *In Proceedings of the 11th International Conference on Parallel and Distributed Systems, IEEE Computer Society*, 2005.
[13] M. Kacimi and K. Yetongnon, "Density-based clustering for similarity search in a p2p network," *In proceedings of the 6th IEEE Symposium on Cluster Computing and the Grid*, 2006.
[14] J. B. Kruskal and M. Wish, "Multidimensional scaling," *In Paper Series on Quantitative Applications in the Social Sciences*, pp. 7–11, 1978.
[15] P. Magdalinos, C. Doulkeridis, and M. Vazirgiannis, "K-landmarks: Distributed dimensionality reduction for clustering quality maintenance," *In Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'06)*, 2006.
[16] L. Wang and L. Khan, "Automatic image annotation and retrieval using weighted feature selection," *In Proceedings of Multimedia Tools and Applications*, vol. 29(1), pp. 55–71, 2006.
[17] R. Yong and T. S. Huang, "Image retrieval: Current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39–62, 1999.