RESEARCH STATEMENT

GOAL AND MOTIVATION

For the past few years, the development in computer science and other areas has made the management of data and information more challenging than ever. New applications such as multimedia technology, bioinformatics and web searching often involve large amount of data in gigabytes or even terabytes. How to efficiently handle these new types of data is an open research problem.

My research goal is to develop new database technologies that lead to efficient and effective querying and knowledge discovering on very large data sets of emerging applications, such as multimedia databases, bioinformatics and web searching, with a focus on similarity queries. The motivation of my research arises from the fact that as the database size increases rapidly, it becomes more and more impractical to use a brutal force approach, such as the linear scan, to query the data. On the other hand, existing access methods may not be suitable for the new data types and/or similarity measures, resulting in a less-optimized performance. Another motivation arises from the fact that similarity queries need to capture the semantic of the queried objects accurately. However, the idea of semantically similar is not the same for different applications. Even for the same application, users may have very different perspectives of similarity. Moreover, as the database size becomes larger, noise level for a given similarity query also increases, leading to query results with lower quality.

RESEARCH EXPERIENCE

My research involves the following areas that are closely related to my research goal.

· **Developing Efficient Indexing Methods for Non-ordered Discrete Data Spaces (NDDS)**

A non-ordered discrete data space (NDDS) is defined as the Cartesian product of domains of non-ordered discrete values. NDDS data is prevalent in applications such as data mining and bioinformatics. A genome sequence database, which consists of elements from a non-ordered discrete domain {*a*, *g*, *t*, *c*}, is a typical application for an NDDS. Unfortunately, existing indexing methods that support similarity queries either cannot be directly applied to an NDDS (e.g., the R-tree) due to lack of essential geometric properties or have less-optimized performance (e.g., the metric trees) due to their generic nature. To solve the problem, we have developed the ND-tree [1, 3] and the NSP-tree [2].

The ND-tree is the first indexing method of its kind to support efficient similarity queries in an NDDS. We extended essential geometric concepts from the continuous data space to an NDDS so that a spatial access method is possible. We have also developed effective tree-construction heuristics that are based on some special characteristics of an NDDS, such as limited alphabet sizes and data distributions.

The NSP-tree is a space-partitioning-based indexing technique for NDDSs, which is different from the data-partitioning-based ND-tree. The NSP-tree improves query performance by ensuring an overlap-free tree structure. We are currently working on indexing techniques for more general applications such as a data space that contains both non-ordered discrete and continuous domains.

REFERENCES

[1] Gang Qian, Qiang Zhu, Qiang Xue and Sakti Pramanik. "The ND-tree: a dynamic indexing technique for multidimensional non-ordered discrete data spaces", in *Proceedings of the 29th International Conference on Very Large Databases* (VLDB 2003), pp. 620-631, Berlin, September 2003.

[2] G. Qian, Q. Zhu, Q. Xue and S. Pramanik. "A Space-Partitioning-Based Indexing Method for Multidimensional Non-Ordered Discrete Data Spaces", in *ACM Transactions on Information Systems* (TOIS) (expected to appear in January 2006).

[3] G. Qian, Q. Zhu, Q. Xue and S. Pramanik. "A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces", in *ACM Transactions on Database Systems* (TODS) (expected to appear in June 2006).

· **Comparing Different Distance Measures for Continuous Data Spaces (CDS)**

A distance measure is an important component of a vector model, which is widely used to support similarity queries. There are numerous distance measures proposed for a CDS. How to choose a proper distance measure for a particular application is an open research issue. We believe that the problem could be solved if the relationships among those different distance measures are understood and effectively utilized. We have theoretically compared two commonly used distance measures, namely, the Euclidean distance (EUD) and the cosine angle distance (CAD). We found that the CAD usually gives a higher rank to vectors with larger variance among its components [1]. We also found that for nearest neighbor queries in high-dimensional data spaces, the query results from the EUD are similar to those of the CAD [2]. The experimental results have corroborated our theoretical analysis. More importantly, we have provided a high-dimensional geometrical model [2] to analyze the relationships among different distance measures. Based on our analysis, we have used the CAD in place of the commonly used EUD in the application of content-based image retrieval (CBIR). Our experimental results have shown that the CAD provides an effective and efficient inter-feature normalization method for the CBIR. We plan to develop a general theoretical framework for comparing other distance measures.

REFERENCES

[1] Gang Qian, Shamik Sural and Sakti Pramanik. "A comparative analysis of two distance measures in color image databases", in *Proceedings of the IEEE 2002 International Conference on Image Processing* (*ICIP 2002*), pp. 401-404, Rochester, September 2002.

[2] Gang Qian, Shamik Sural, Yuelong Gu and Sakti Pramanik. "Similarity between Euclidean and cosine angle distance for nearest neighbor queries", to appear in *Proceedings of the 19th Annual ACM Symposium on Applied Computing* (*SAC 2004*), Nicosia, Cyprus, March 2004.

· **Generating Effective Color Features using the HSV Color Space**

Color is an important component for similarity queries in image databases. RGB is the most commonly used color space to generate color features. In this research, we have done in-depth analysis of the visual properties of the HSV color space. In particular, we have studied histogram generation [1] and image segmentation applications [2] using the HSV color space. The HSV color space is fundamentally different from the RGB color space since it separates out the intensity (luminance) from the color information (chromaticity). Of the two chromaticity axes, we found that a small change in Hue is visually more detectable than that of the Saturation. Based on this special property of the HSV color space, we developed a novel histogram generation technique where a perceptually smooth transition of color is obtained in the feature vector. This enables us to use a window-based comparison of histograms so that similar colors can be matched between a query and each image in the database. We also segmented color

images, using features extracted from the HSV color space, as a step in an object-based matching approach to CBIR. We were able to determine the position of objects so that images may be compared at the object level. Both image segmentation and color histogram based image retrieval show better result compared to similar approaches based on the RGB color space.

REFERENCES

[1] Shamik Sural, Gang Qian and Sakti Pramanik. "A histogram with perceptually smooth color transition for image retrieval", in *Proceedings of the 4th International Conference on Computer Vision, Pattern Recognition and Image Processing* (*CVPRIP 2002*), pp. 664-667, Durham, March 2002.

[2] Shamik Sural, Gang Qian and Sakti Pramanik. "Segmentation and histogram generation using the HSV color space for content based image retrieval", in *Proceedings of the IEEE 2002 International Conference on Image Processing* (*ICIP 2002*), pp. 589-592, Rochester, September 2002.

· **Developing Hybrid RAM/Disk-based Index Techniques for Large Text Databases**

Most indexing methods for text documents, such as the Tries, are RAM-based approaches. To support similarity queries for very large text databases such as the web, there might not be enough RAM to hold the whole index structure. On the other hand, pure disk-based indexing methods such as the String B-tree do not fully utilize the large amount of RAM available these days. Therefore, we believe that an efficient hybrid index structure that combines the benefits of both the RAM-based and the disk-based approaches will be a good solution to the problem [1, 2]. Our experimental results have shown that a hybrid index structure is quite promising for supporting similarity queries on very large text databases.

REFERENCES

[1] Q. Xue, S. Pramanik, G. Qian and Q. Zhu. "The Hybrid Digital Tree: a new indexing technique for large string databases", in Proceedings of the 7th International Conference on Enterprise Information Systems (ICEIS 2005), vol. 1, pp. 115-121, Miami, May 2005.

[2] Q. Xue, S. Pramanik, G. Qian and Q. Zhu. "A hybrid index structure for querying large string databases", in *International Journal of E-Business*, vol. 3, no. 3/4, pp. 243-254, 2005.

· **Analyzing the Filtering Step for Genome Sequence Searching**

A common approach to search a genome sequence database is to first select candidate regions from the database, a step called filtering. Local alignments are then conducted on the candidate regions to find true homologous (similar) regions. Since local alignments are very expensive, the searching time could be significantly reduced if the filtering step provides a small and accurate set of candidate regions. Exact matching of q-grams is a widely used method for the filtering step in popular systems such as the BLAST. However, this approach may result in a large set of candidate regions with low accuracy, as there could be too many hits in the database. We are conducting both experimental and theoretical analysis on the behavior of the filtering step while allowing some mismatches for the q-grams [1]. Our initial experimental results have shown oblivious improvement over the existing methods.

REFERENCES

[1] "Approximate q-gram Matching in Genome Sequence Databases", in preparation for publication.

FUTURE PLANS

In the near future, I plan to investigate new indexing techniques and feature generation methods to support

efficient similarity queries in emerging applications that involve a large amount of non-traditional data. Different applications usually have their own specific challenges; for example, how to generate semantically meaningful features for different multimedia objects like movies and music, how to index biological information using the edit distance, and how to efficiently support web queries with different semantic meanings. All these important questions require careful analysis of their underlying principles. Since these are all promising new areas with a wide range of applications, there are a lot of collaboration and funding opportunities from the industry and government agencies like NIH, NSF and DARPA. I would also like to develop collaborative research with my colleagues in these areas.

STUDENT INVOLVEMENT

I would expect that my graduate students would become actively involved in my research projects. While any project of my research plan can be a major research agenda or theses, there is nothing inherent in it that precludes undergraduate students from making significant contributions.