Ramakrishna Varadarajan
ramkris@cs.wisc.edu
http://pages.cs.wisc.edu/~ramkris

Research Statement

Overview

Graphs provide a powerful primitive for modeling data in a variety of applications. Nodes in graphs usually represent real world objects and edges indicate relationships between objects. Examples of data modeled as graphs include social networks, biological networks, and dynamic network traffic graphs. The problem of effective search, retrieval and pattern mining from such graphs has been receiving much attention recently. World Wide Web can be naturally viewed as a graph of pages and hyperlinks. Likewise, relational and XML databases can be viewed as graphs of hyperlinked entities. *Keyword search querying* has emerged as one of the most effective paradigms for information discovery, especially over HTML documents in the World Wide Web. One of the key advantages of keyword search querying is its simplicity – users do not have to learn a complex query language, and can issue queries without any prior knowledge about the structure of the underlying data. Recently, the problem of keyword search over relational and XML databases has received much attention. Frequent pattern mining has been a focused theme in data mining research for over a decade. Among the various kinds of graph patterns, *frequent substructures* are very basic patterns that can be discovered in a collection of graphs. Recent studies have developed several frequent substructure-mining methods.

Past Research (*Ranked Keyword Search on Graph-structured databases*)

My PhD thesis work [1] developed techniques for user-friendly, high quality and efficient searching of interlinked (graph-structured) databases. Several ranked search methods on data graphs have been studied in the recent years. Given a top-*k* keyword search query on a graph and ranking criteria, a *keyword proximity search* finds how the query keywords are related in the graph through semantic connections by computing top-*k* answers (a substructure of the graph) containing all query keywords. Current Web search engines return a list of individual web pages ranked by their relevance to the query. The limitation of this approach is that it operates at the page level, which ignores the specific context where the keywords are found in the pages. This degrades the quality of search results, especially for long uncorrelated queries (in which individual keywords rarely occur together in one document), where a single page is unlikely to satisfy the user's information need. It is cumbersome for the user to locate the most desirable text fragments relevant to the query due to the amount of data in

each page and large number of interconnected pages. My research [4, 8, 10, 11] tackles this problem by applying keyword proximity search on the web and the document graph of web documents to find top-*k* answers that satisfy user's information need and increase user satisfaction.

Another effective ranking mechanism applied on data graphs is the authority flow based ranking mechanism. Given a top-*k* keyword search query on a graph, an *authority-flow based search* [5, 7] finds the top-*k* answers where each answer is a node in the graph ranked according to its relevance and importance to the query. This technique was first applied on the web and later over databases and XML. My research developed techniques to improve the authority flow based search on data graphs by creating a framework to explain and reformulate them taking in to consideration user feedback and preferences.

Current Research (*Pattern Mining in Graph-structured Databases*)

A common mining task on graph databases is to find frequent common (sub) graphs in the database. This important problem has attracted many research proposals. Existing methods are designed to take as input a set of graphs and an additional input parameter, such as the minimum support. These algorithms then output all sub-graph patterns that are present in the database with a frequency that is greater than or equal to the minimum support. In practice, given the exploratory setting in which *frequent graph mining* is used, picking this minimum support input parameter value is challenging for the end-user, even when the user is intimately familiar with the actual database that s/he is using [14]. Consequently, users have to resort to making a series of guesses about the minimum support. A lower minimum support parameter value can make the mining process slow (sometimes intolerably slow if the graph dataset is very large), and produces many results that the user may not be interested in. On the other hand, a higher minimum support value can result in very few results, thereby potentially missing many interesting patterns. Users have adapted to this limitation by selecting and reselecting "appropriate" minimum support based on a trial and error approach.

Another practical problem [14] with existing mining methods is that the results typically get outputted to the user in large bursts towards the end of the mining process. Given the high complexity of frequent (sub) graph mining methods, this means that the user has to wait for a long time, especially when analyzing large datasets, to view *any* results. Since mining is an exploratory task that is often used in an interactive setting, a more user-friendly paradigm is to have an online algorithm that quickly shows the initial results, and then continually keeps adding more results. Thus, what is desired in practice is an efficient, parameter-free, online graph mining tool that the user can simply point the data to, and quickly start seeing results presented progressively in *decreasing support order*.

Apart from frequent pattern mining in graphs, *mining significant subgraphs* has attracted much interest recently. Even when lot of progress has recently been made in graph pattern mining, mining significant patterns still remains a challenging task because the anti-monotone property which is at the core of powerful pruning techniques in data mining cannot be applied. This makes the mining slower and scalable techniques are needed for optimal substructure mining for various objective functions.

Future Research (*Information Discovery on Domain Data Graphs & Significant Substructure Mining*)

My short term research goal is to develop effective Information Discovery techniques on healthcare/clinical [2, 3, 6], biological [9, 12] and intellectual property [13] databases by applying sophisticated ranked search techniques taking into consideration the domain characteristics. In particular, an increasing amount of data is stored in biological sources, like Entrez Gene, PubMed, and OMIM. Entities of the sources are interconnected through semantic links, created manually or automatically (e.g., using BLAST). Incorporating domain knowledge – physician vs. biologist profiles, wisdom of domain experts, precision vs. recall requirements and preferred schema for the output – in an effective and efficient way is challenging, given the often large size of such data collections. Regarding clinical databases, in addition to the above challenges, we need to address privacy and data heterogeneity – no standard format for medical record representation – issues.

My long term research goal is to build a search engine for querying generic domain databases, which can be customized for different domains and users. A language to describe the domain intricacies is needed. User feedback must also be captured to personalize the search experience. Another interesting long term research goal is addressing the scalability issue in significant substructure graph mining tasks, while still maintaining the quality of the mined patterns.

REFERENCES

1. *Ramakrishna Varadarajan*: "*Ranked Keyword Search on Graph-Structured Databases*: Techniques for User-friendly, High Quality and Efficient Information Discovery on Data Graphs". Publisher: VDM Verlag (February 24, 2010). ISBN-10: 3639237269. ISBN-13: 978-3639237269.
2. *Ramakrishna Varadarajan*, Vagelis Hristidis and Fernando Farfan: "*Searching Electronic Health Records*". Book Details: "Information Discovery on Electronic Health Records". CRC - Taylor & Francis, December 2009. Editor: Vagelis Hristidis.
3. Fernando Farfan, *Ramakrishna Varadarajan* and Vagelis Hristidis: "*Electronic Health Records*". Book Details: "Information Discovery on Electronic Health Records". CRC - Taylor & Francis, December 2009. Editor: Vagelis Hristidis.

4. *Ramakrishna Varadarajan*, Vagelis Hristidis and Tao Li: "*Beyond Single-Page Web Search Results*", IEEE Transactions on Knowledge and Data Engineering (TKDE) 2008.
5. Vagelis Hristidis, Yannis Papakonstantinou and *Ramakrishna Varadarajan*: "*Using Proximity Search to Estimate Authority Flow*", IEEE Transactions on Knowledge and Data Engineering (TKDE) 2010.
6. Vagelis Hristidis, *Ramakrishna Varadarajan*, Paul Biondich, Redmond Burke and Michael Weiner: "*Information Discovery on Electronic Medical Records Using Authority-Flow Techniques*". In BMC Medical Informatics and Decision Making, 2010.
7. *Ramakrishna Varadarajan*, Vagelis Hristidis and Louiqa Raschid: "*Explaining and Reformulating Authority Flow Queries*" (full paper), IEEE 24th International Conference on Data Engineering (ICDE) 2008, Cancun, Mexico. (*Acceptance rate* – 19% *Impact factor* – 0.97).
8. *Ramakrishna Varadarajan* and Vagelis Hristidis: "*A System for Query-specific Document Summarization*" (full paper), ACM 15th Conference on Information and Knowledge Management (CIKM) 2006, Arlington, VA, pages 622-631. (*Acceptance rate* – 15% *Impact factor* – 0.90).
9. *Ramakrishna Varadarajan*, Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis lbanez and Hector Rodriguez-Drumond: "*Flexible and Efficient Querying and Ranking on Hyperlinked Data Sources*" (full paper), Extending Database Technology (EDBT) 2009, Saint-Petersburg, Russia. (*Acceptance rate* – 32.50% *Impact factor* – 0.90).
10. *Ramakrishna Varadarajan*, Vagelis Hristidis and Tao Li: "*Searching the Web using Composed Pages*" (poster paper), ACM SIGIR Conference on Research and Development on Information Retrieval 2006, Seattle, WA, pages 713-714. (*Acceptance rate* – 37% *Impact factor* – 0.94).
11. *Ramakrishna Varadarajan* and Vagelis Hristidis: "*Structure-Based Query Specific Document Summarization*" (poster paper), ACM 14th Conference on Information and Knowledge Management (CIKM) 2005, Bremen, Germany.
12. *Ramakrishna Varadarajan*, Felix Eichinger, Jignesh Patel, Matthias Kretzler: "*Molecular Re-Classification of Renal Disease using Approximate Graph Matching, Clustering and Pattern Mining*" (poster paper), ISMB 2010, Boston.
13. Vagelis Hristidis, Eduardo Ruiz, Alejandro Hernandez, Fernando Farfan, *Ramakrishna Varadarajan*: "*PatentsSearcher*: *A Novel Portal to Search and Explore Patents*" (http://www.patentssearcher.com/). In 3rd International Workshop on Patent Information Retrieval (PaIR 2010), ACM CIKM 2010.
14. *Ramakrishna Varadarajan*, Jignesh Patel: "*Practical and Efficient Online Frequent Graph Mining*". (Under review), in PVLDB.
15. *Ramakrishna Varadarajan*, Vagelis Hristidis, Fernando Farfan: "*Comparing Top-k XML Lists*" (Under review), in EDBT 2011.
16. Vijil Chenthamarakshan, *Ramakrishna Varadarajan*, Prasad Deshpande and Raghuram Krishnapuram: "*WYSIWYE: An Algebra for Expressing Spatial and Textual Rules for Information Extraction*". (Under review).