8ª MOSTRA ACADÊMICA UNIMEP
26 a 28/10 de 2010
Tema: Desafios da Educação Superior na Agenda do Novo Milênio

8º Congresso de Pós-Graduação

MINERAÇÃO BASEADA EM GRAFOS APLICADA À ÁREA BIOMÉDICA

Autor(es)

RODRIGO DE SOUSA GOMIDE

Co-Autor(es)

MARINA TERESA PIRES VIEIRA

Orientador(es)

MARINA TERESA PIRES VIEIRA

1. Introdução

A Mineração de Dados é uma fase do processo de Descoberta de Conhecimento em Bases de Dados (KDD) usada para encontrar regras e padrões em um conjunto de dados. Conforme Han e Kamber (2006) a mineração de dados é definido como sendo o ato de extrair ou “minerar” conhecimento de uma grande quantidade de dados.

Algumas vezes, dados do mundo real se apresentam, devido à própria estrutura em que foram organizados, dispostos de forma complexa, de tal modo que sua representação em formato tabular e/ou gráfico dificultaria uma compreensão analítica.

A aplicação de algoritmos de mineração de dados em estruturas complexas mal organizadas pode comprometer a potencialidade do processamento desses algoritmos. Esses algoritmos algumas vezes podem não encontrar regras realmente relevantes devido à organização relacional dos dados.

De acordo com Cook (2007), banco de dados relacionais e lógica de primeira ordem são duas representações populares, porém não são completamente suportadas no processo de mineração de dados. A mudança de estrutura, modelo relacional para grafos, pode oferecer um melhor conjunto de padrões após o processo de mineração de dados.

Pode-se definir a mineração de grafos como o processo de busca do conhecimento de subestruturas através de um conjunto de estruturas representadas por meio de grafos. Essa subárea da mineração de dados utiliza as mesmas tarefas, porém usando algoritmos e regras adaptados para representação de grafos.

A mineração de grafos tem sido aplicada em diversas áreas do conhecimento. Takizawa, Yoshida e Katoh (2007) propuseram encontrar relações nas estruturas de salas de apartamentos em aluguel através da extração de subestruturas de grafos, cuja motivação é prever modelos de apartamentos, com maior precisão.
Song e Chen (2006) apresentam alguns algoritmos de mineração em grafos que podem ser aplicados em redes regulatórias genéticas. Também Borgelt e Berthold (2002) propuseram um algoritmo usado na busca de fragmentos em um conjunto de moléculas dispostos em grafos; a busca ajuda a identificar diferentes tipos de grupos de moléculas, permitindo encontrar doenças como câncer e vírus como HIV.

Existem trabalhos com aplicações na área da ciência da computação existem trabalhos como o de Shrivastava e Pal (2009) que propõem a construção de um framework que abranja as três etapas do processo de KDD voltado à mineração de grafos, iniciando com a construção de grafos durante o pré-processamento, seguindo com a descoberta de sub-grafos freqüentes usando os algoritmos mais populares e concluindo com a visualização dos grafos no pós-processamento. Lam e Chan (2008) também propõem um novo algoritmo de mineração de grafos aplicado à busca de padrões em estruturas de layout de páginas da WEB.

Este trabalho pretende usar a mineração de grafos para tratar um problema da área biomédica. Esse problema surgiu devido a um projeto de pesquisa em andamento, desenvolvido por um grupo de pesquisa liderado por docentes da UFSCar, USP e UNIMEP, que busca desenvolver recursos para analisar dados sobre a doença Anemia Falciforme.

2. Objetivos

O objetivo deste trabalho é usar os conceitos da mineração de grafos para propor uma forma de modelar e minerar dados da área biomédica.

Especificamente, pretende-se propor uma representação na forma de grafos de um conjunto de dados distribuídos em várias tabelas de uma base de dados relacional da área biomédica, para que esses dados possam ser processados por um algoritmo de mineração de grafos.

Devido às características intrínsecas do problema a ser tratado, será necessário adaptar um algoritmo existente.

3. Desenvolvimento

A Mineração de Texto aplicada à área biomédica se tornou um projeto de pesquisa envolvendo diversas instituições, Universidade de São Paulo Unidade de Ribeirão Preto (USP), Universidade Federal de São Carlos (UFSCAR), Universidade de São Paulo Unidade de São Carlos e Universidade Metodista de Piracicaba (UNIMEP).

Recentemente um dos pesquisadores deste grupo, mestrando pela UNIMEP, pesquisou em seu projeto um método de se aplicar a Mineração de Dados Multi-relacional na área biomédica. Nesse trabalho o autor pesquisa a praticidade da utilização da mineração multi-relacional em encontrar padrões numa base de dados de experimentos da área biomédica, propondo a implementação de um algoritmo adaptado a área em questão.

Atualmente os dados sobre Anemia Falciforme são extraídos de artigos científicos da área médica, que relatam resultados de tratamentos em pacientes dessa doença. Esses experimentos apresentam características, tais como: o tratamento usado para combater a doença; os efeitos colaterais do tratamento e as complicações causadas pela doença; os benefícios relativos ao tratamento; e a quantidade de pessoas envolvidas nas experiências. Por exemplo, um portador da anemia falciforme pode apresentar as seguintes complicações: crise de dor recorrente e síndrome torácica aguda. O tratamento com hidroxiuréia (hu) pode amenizar essas complicações, porém causar o seguinte efeito colateral: anemia aguda (GULBIS et. al., 2005).

Baseado nessas premissas foi possível montar o esquema Entidade-Relacionamento compatível com a realidade atribuída. Recentemente um novo problema relativo ao assunto foi identificado por uma especialista da área médica integrante do projeto Anemia Falciforme. Sabe-se que os tratamentos relacionados à Anemia Falciforme apresentam algumas vezes efeitos colaterais e/ou complicações. Esses efeitos colaterais e/ou complicações podem ser solucionados com outros métodos de tratamento.

As novas doenças, acarretadas pelos efeitos colaterais de um tratamento, também devem ser consideradas, ou seja, a base de dados além de conter artigos que orientam o tratamento da Anemia Falciforme também devem conter artigos que contemplem o tratamento dessas novas doenças.

De acordo com essa nova perspectiva se deseja cruzar as informações relativas à Anemia Falciforme com os dados de outros agravantes de saúde. Especificamente um dos objetivos do projeto é cruzar os tratamentos, benefício, efeitos colaterais e complicações de artigos que proponham tratamentos alternativos a efeitos colaterais da Anemia Falciforme.

O desenvolvimento do projeto usa como métodos de pesquisa a estrutura de índice de adjacência na organização dos dados (WANG et. al., 2004). Para extração das informações adota-se o algoritmo de mineração de grafos gSpan (HAN e YAN, 2002).

Ambos os métodos que implementam esses recuros serão alterados para que possam atender as necessidades exigidas pela da área biomédica.

A) ÍNDICE DE ADJACÊNCIA

O Índice de Adjacência foi um método desenvolvido por Wang et al. (2004), e seu objetivo é criar uma forma de indexação que auxilie no desempenho da mineração de dados baseada em grafos, em bases de dados acessadas em larga escala no disco.

O ADI é uma estrutura de três níveis: índice para arestas; identificação dos grafos no qual as arestas estão contidas; e informação de adjacência, conforme apresentado na figura 1.

Na figura 1, o primeiro nível corresponde ao índice das arestas conhecido também como tabela aresta. Cada aresta aponta para um conjunto de grafos presente no nível intermediário, este nível é responsável pela identificação dos grafos. Por fim, o último nível, é responsável por informar a estrutura do grafo, e é conhecida como informação de adjacência.

B) GSPAN

Criado por Han e Yan (2002), o gSpan foi uma alternativa encontrada para resolver a questão de extração de informação em um conjunto de grafos. Esse algoritmo é fundamentado pela abordagem Pattern Growth. Ele foi projetado para não realizar a busca em grafos já encontrados previamente, evitando sua duplicação. Mesmo assim garante uma busca completa dos grafos mais freqüentes.

Esse algoritmo faz uso da busca em profunidade (DFS) para encontrar os padrões de grafos freqüentes.

4. Resultado e Discussão

O desafio proposto é adaptar os métodos discutidos anteriormente de tal forma que atenda as exigências solicitadas dentro da área biomédica.

O número de pacientes aos quais foram aplicados experimentos, por exemplo, é uma informação valiosa dentro desta área. O suporte dos algoritmos de mineração de dados geralmente são definidos através de uma porcentagem no qual são usados na busca por padrões.

Na área biomédica, o valor percentual caracterizado no suporte mínimo passa a não ter significância. Um atributo relevante a área em questão, diz respeito ao número de pacientes. Existe então a necessidade de adaptar o número de pacientes ao suporte.

Para resolver este problema, o número de pacientes influenciará diretamente o segundo nível da estrutura ADI. O rótulo das arestas indica os pacientes envolvidos.

Para um grafo que possuísse, por exemplo, aresta (A, 5, B) temos a seguinte interpretação: cinco pacientes que fizeram o tratamento A acarretaram a um complicação B. Supondo que esta aresta estivesse presente em um grafo G1, no segundo nível da estrutura ADI o id G1 se repetiria cinco vezes.

A figura 2 apresenta um dataset carregado pela aplicação. Cada grafo contido no dataset, pode ser visualizado.

Após o carregamento do conjunto de grafos, a estrutura ADI é construída conforme mostra a figura 3.

5. Considerações Finais

O trabalho apresentando encontra-se em processo de desenvolvimento, mas sabe-se que, com a estrutura apresentada, será possível extrair padrões freqüentes de tal forma que aponte soluções alternativas de tratamentos a doenças secundárias causadas pela terapia da Anemia Falciforme.

Também será possível extrair doenças similares entre tratamentos distintos. Enfim, o processo de mineração de grafos aplicado à estrutura proposta visa adquirir informações valiosas a respeito do cruzamento de experimentos voltados para Anemia Falciforme e suas ramificações.

Referências Bibliográficas

BORGELT, Christian; BERTHOLD, Michael. Mining molecular fragments: finding relevant substructures of molecules. The 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, 2002, page 51

CHI, Yun; et. al. Mining Closed and Maximal Frequent Subtrees from Databases of Labeled Rooted Trees. IEEE Transactions on Knowledge and Data Engineering, February 2005, page 190 – 202

COOK, Diane J. HOLDER, Lawrence B. Mining Graph Data. Wiley. 2007

GULBIS, Béatrice; et. al. Hydroxyurea for sickle cell disease in children and for prevention of cerebrovascular events: the Belgian experience. The American Society of Hematology, April 2005, page 2685 – 2690

HAN, Jiawei; YAN, Xifeng. gSpan: Graph-Based Substructure Pattern Mining. The 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, page 721 – 724, 2002

HAN, Jiawei; YAN, Xifeng. CloseGraph: Mining Closed Frequent Graph Patterns. In: Proceedings of the 2003 Conference on Knowledge Discovery and Data Mining (SIGKDD2003), 2003, page 286 – 295

LAM, Winnie; CHAN, Keith; Analyzing Web Layout Structures using Graph Mining. IEEE International Conference on Granular Computing, Hangzhou, China, August 2008, page 361 – 366

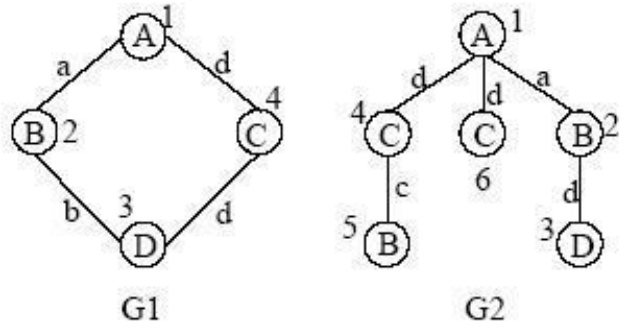
SHRIVASTAVA, Swapnil; N. PAL, Supriya; Graph mining framework for finding and visualizing substructures using graph database. Advances in Social Network Analysis and Mining (ASONAM 2009), Athens, Greece, July 2009, page 379-380

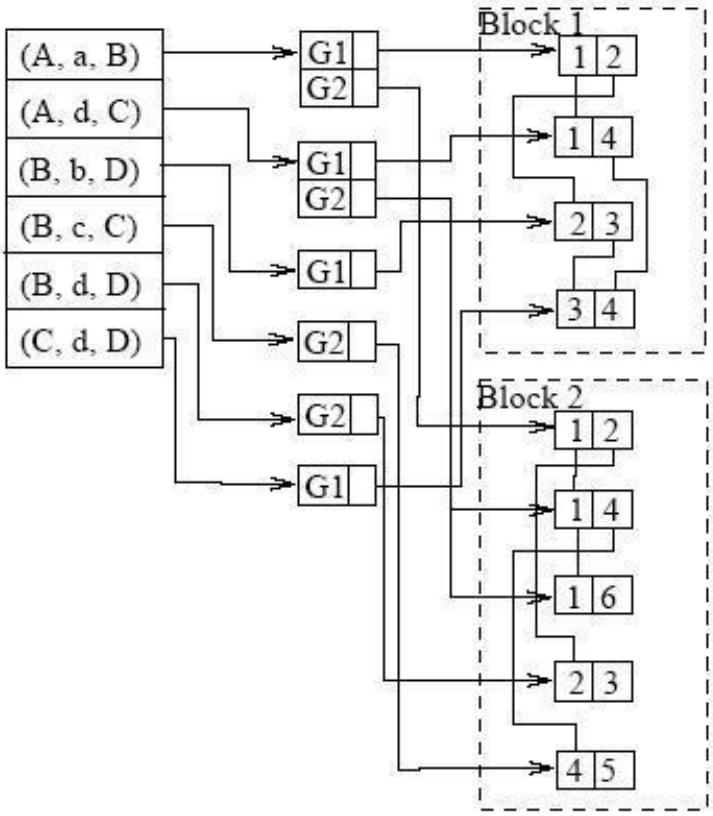
SONG, Yongling; CHEN, Su-Shing; Item set based graph mining algorithm and application in genetic regulatory networks. IEEE International Conference on Granular Computing, Atlanta, USA, May 2006, page 337 – 340

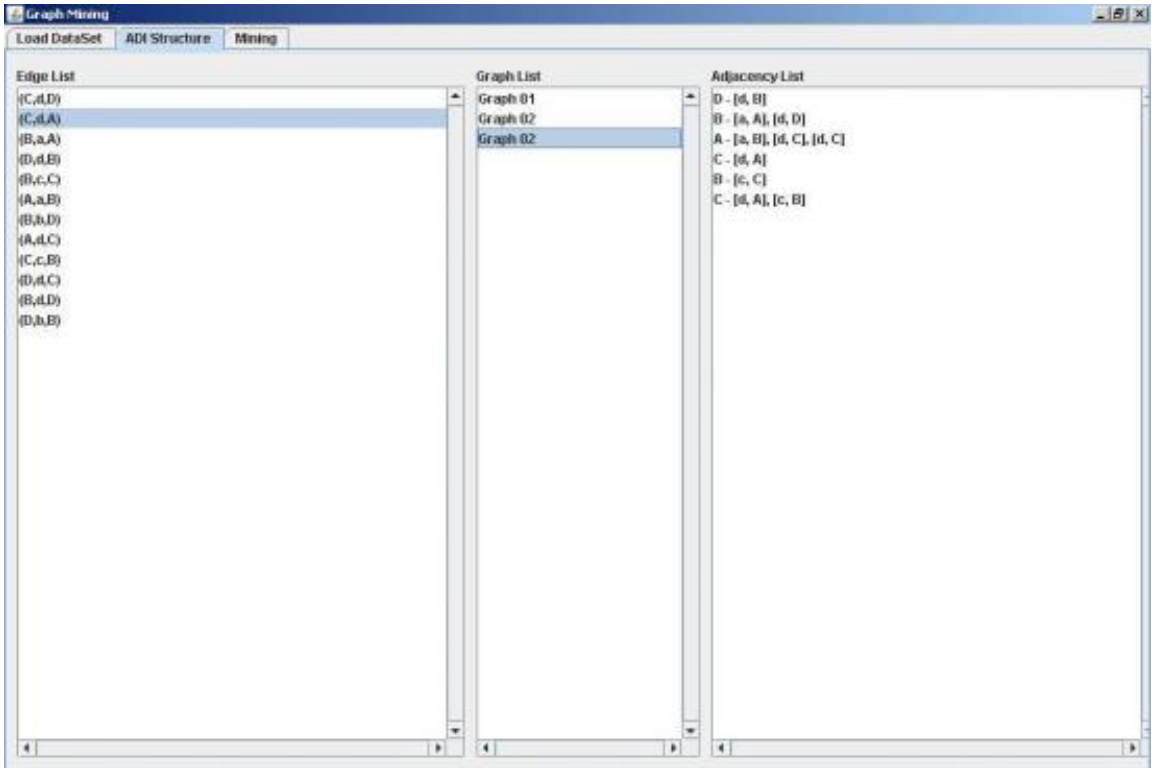
TAKIZAWA, Atsushi; YOSHIDA, Kazuma; KATOH, Naoki; Applying graph mining to discover substructures of room layouts which affect the rent of apartments. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Montréal, Canada, October 2007, page 3512 – 3518

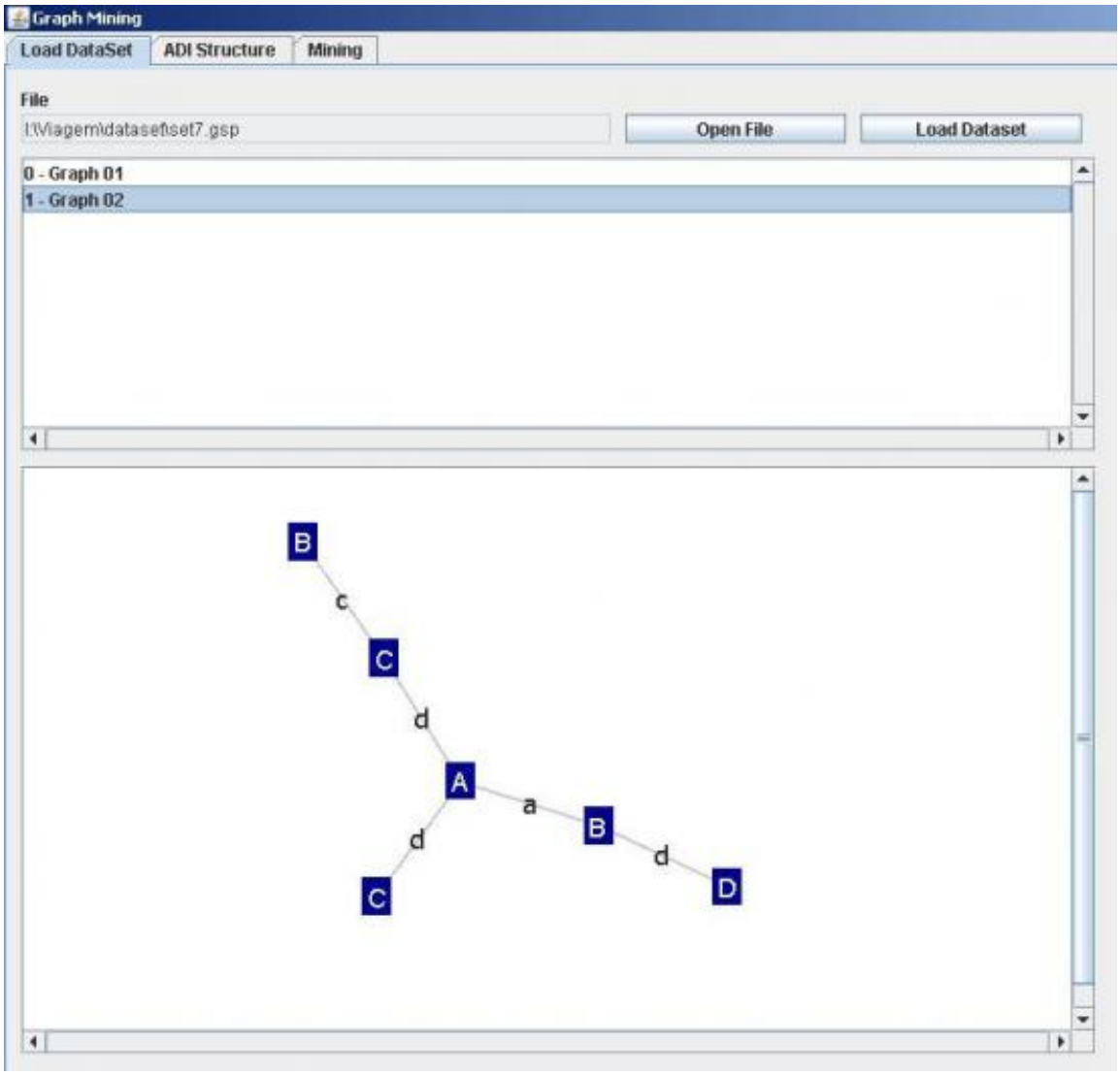
WANG, Chen; et.al; Scalable mining of large disk-based graph databases. In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD `04), Washington, U.S.A., August 2004, page 316-325

Anexos


A
B
C
D
a
b
d
d
1
2
3
4
G1
A
C
C
B
B
D
d
d
a
c
d
1
4
6
2
3
5
G2


Edges
Graph−ids (on disk)
Adjacency (on disk)
(A, a, B)
(A, d, C)
(B, b, D)
(B, c, C)
(B, d, D)
(C, d, D)
G1
G2
Block 1
Block 2
1 2
1 4
2 3
3 4
1 2
1 4
1 6
2 3
4 5


Graph Mining
Load DataSet
ADI Structure
Mining
Edge List
(C,d,D)
(C,d,A)
(B,a,A)
(D,d,B)
(B,c,C)
(A,a,B)
(B,b,D)
(A,d,C)
(C,c,B)
(D,d,C)
(B,d,D)
(D,b,B)
Graph List
Graph 01
Graph 02
Graph 02
Adjacency List
D - [d, B]
B - [a, A], [d, D]
A - [a, B], [d, C], [d, C]
C - [d, A]
B - [c, C]
C - [d, A], [c, B]


Graph Mining
Load DataSet
ADI Structure
Mining
File
I:\Viagem\dataset\set7.gsp
Open File
Load Dataset
0 - Graph 01
1 - Graph 02
B
c
C
d
A
a
B
d
d
C
D