An in-depth study of the syntactic analysis of graphs using search matching algorithms, to the degree of sub graphs

Student number:
email:
Supervisor: Dr
Dept Computer Science, University of Pretoria

August 20, 2015

1 Graph Basics

1.1 Graph Overview

A Graph in Mathematics and Computer Science is defined as a pair $G = (V, E)$, where V is the set of vertices and E is the set of edges, formed by pairs of vertices with each other. Figure 1 demonstrates the structural attributes of a simple graph.

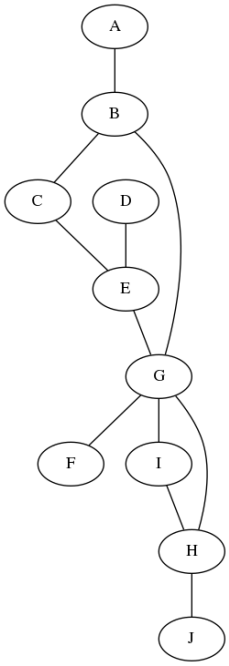


Figure 1: Representation of a graph

Graphs can either directed or they can undirected. This means that the edges in the graph could have an abscence of direction as in the 1 above, or they could have a direction showing from which vertice the edge is coming from, and to which vertice the edge is going to. Figure 2 demonstrates a directed graph, commonly known as a digraph. This characteristic is demonstrated by the edge 3, that goes from node B to node G, and also by edge 2 that goes from node B to node C.

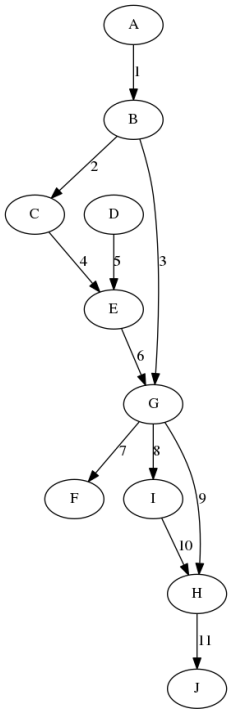


Figure 2: Representation of a digraph

Note that the undirected graph mentioned above is commonly referred to as a bigraph because the direction of its edges could be percieved as to going in both direction as it is not specifed.

In this paper, the focus is primarily on directed graphs, and they shall be referred to as digraphs from here onward.

1.2 Digraph

A Digraph in mathematics is defined as is a pair of disjoint vertices and edges (V, E), and their repective mappings that comprises of two components, namely the initial vertix and terminal vertice of each edge i.e. each edge has a initial vertix:

$$Ei \to Vi \tag{1}$$

and a terminal vertix:

$$Vj \to Ej \tag{2}$$

for some vertices Vi,Vj in V and edges Ei,Ej in E [7] refer to the figure 2.

1.3 Graph representation

Graphs are represented in a variaty of ways, from adjacency lists, incident matrices and adjacency matrices. The algorithms that are studied in this paper make us of adjacency matrices and adjacency list representations of graphs.

1.3.1 Adjacency matrices

An adjacency matrices is a nxn matrix A, with $A(i,j) = 1\ iff(i,j)E$ [9]. This means that wherever there is an edge in the graph, it is denoted by a 1 in the matrix, places in the matrix where there is an absence of an edge, are denoted by 0.

Figure 3 depicts the association between the graph in figure 1 and its adjacency matrix.

	A	B	C	D	E	G	F	I	H	J
A	0	1	0	0	0	0	0	0	0	0
B	1	0	1	0	0	1	0	0	0	0
C	0	1	0	0	1	0	0	0	0	0
D	0	0	0	0	1	0	0	0	0	0
E	0	0	1	1	0	1	0	0	0	0
G	0	1	0	0	1	0	1	1	1	0
F	0	0	0	0	0	1	0	0	0	0
I	0	0	0	0	0	1	0	0	1	0
H	0	0	0	0	0	1	0	1	0	1
J	0	0	0	0	0	0	0	0	1	0

Figure 3: Representation of a graph and its associated adjacency matrix

1.3.2 Adjacency list

An Adjacency list is vertices of a graph, of which each vertice is connected to the list. The vertices in an adjacency list point to their own list of edges that they are connected to(i.e. the list contains the edges that connect them to other vertices).

Figure 4 depicts the association between the graph in figure 1 and its adjacency list.

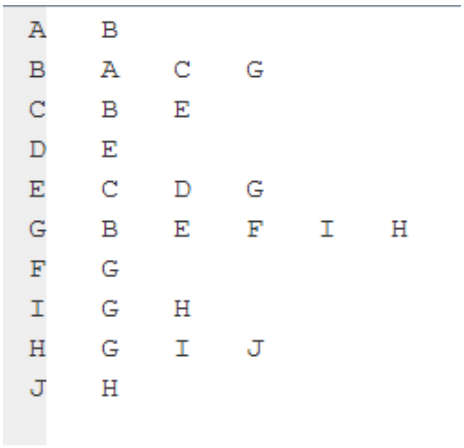


Figure 4: Representation of a graph and its associated adjacency list

1.4 Supergraphs and subgraphs

Let G_A be be graph define as follows $G_A = (V_A, E_A)$ and let G_B be another graph that is defined as follows $G_B = (V_B, E_B)$ where V_A,V_B are sets of vertices and E_A,E_B are sets of edges. In graph theory, a graph G_A is said to be a subgraph of graph G_B, and graph G_B is said to be a supergraph of graph G_A if all the vertices and edges that are in graph G_A are also in graph G_B, that is [3]:

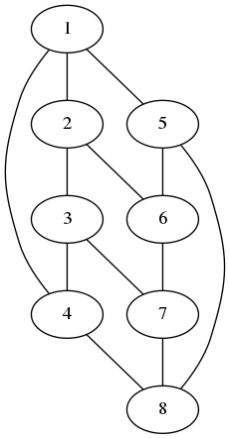
(1) $V_A \subseteq V_B$, and

(2) Every edge of G_A is also an edge in G_B.

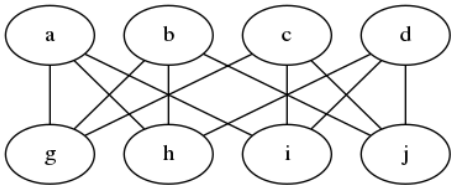
In figure 1, the graph constructed by vertices D, E, G, F and I is the sub-graph of the entire graph. And thus the graph is a super-graph of the sub-graph constructed by the vertices, namely D, E, G, F and I.

1.5 Graph Isomorphism

Two graphs are said to be isomorphic if they are syntactically similar to each other iff there is a bijection between their respective nodes which make each edge of G_A correspond to exactly one edge of G_B, and vice versa [12], i.e. the graphs are structurally the same to each other. This property is demonstrated in figures 5b.



(a) Isomorphic graph on the left



(b) Isomorphic graph on the right

Figure 5: A demostration of two isomorphic graphs

The two graphs look very different, but when they are further inspected, it is evident that the two are a representation of the same structural scheme or maybe even the same graph that has been rearranged.

Consider the vertice 1 from the left-most graph, it has three edges going to and from vertices $2, 4$ and 5, and now consider the vertix a from the right-most graph, it also has three vertices going to and from it, namely g, h and i. These two vertices have the same structural, but from two different graphs, the same goes for all the other vertices in both graphs .i.e each vertice in the left-most graph can be associtated with one in the right-most graph as we did for vertices 1 and a.

2 Search Matching Algorithm

2.1 Introduction

Subgraph ismorphis can be determined using a brute-force approach on the tree representation of a graph G_A. Though this technique is effective, it is however not efficient, this is because all the possible permutation subgraphs of a graph G_A are tested against the graph G_B to determine if there are subgraphs in graph G_A that are isomorphic to graph G_B.The number of subgraphs of a graph G_A increase at an exponential rate with every addition of a vertice V_n into the graph, thus the total number of subgraphs that can be evaluated are

$$ST = 2^{n/2} \tag{3}$$

where ST is the total number of subgraphs and n the number if vertices in the graph G_A.

The matching process is computationally expensive due to this very fact, that is the more vertices there are in the graph G_A, the more expensive it becomes to detect the subgraph ismorphisms because of the amount of subgraphs it has and thus must be evaluated.

This paper explores two graphs that are very effective with regards to graph isomorphis and subgraph isomorphism detection. The algorithms that are invesicated are the Ullman Algorithm and the VF2 algorithm.

2.2 Ullman Algorithm

2.2.1 Algorithm

The Ullman algorithm was developed by J.R.Ullman and was published in his paper titled "An Algorithm for Subgraph Isomorphism" [1].The algorithm performs graph matching on an adjacency matrix representation of both the graphs, and uses the depth search first(DSF) recursicve approach to traverse through the graphs and perform the graph matching process. The Ullman algorithm improves the effiency of the brute-force approach at detecting subgraph ismorphisms by deductively eleminating nodes in the tree that are in graph G_A, but are not in graph G_B, thus reducing the number of subgraphs that are matched against graph G_B to determine ismorphism.

The algorithm starts by building a starting adjacency matrix $M0$ using the two adjacency matrix representations of graphs G_A and G_B using the following procedure.

(1) Construct a $n * m$ matrix where n is the number of rows of graph G_B and m is the number of colums of graph G_A.

(2) Set all the entries in the matrix to the value of 1.

(3) Apply the following rule: Set the values in $M0$ to 0 for all $M0_{ij}$ where the degree of a vertice in graph G_A at j is greater then the degree of the same vertice in graph G_B .i.e.

$$deg(Ai) < deg(Bj). \tag{4}$$

A more formal representation of this rule is as follows

$$f(x) = \begin{cases} 1, & \text{if } deg(Ai) \geq deg(Bj) \\ 0, & \text{otherwise } \forall \text{i,j} \end{cases}$$

When the starting matrix has been constructed, we systematically permute matrices M^d from the starting matrix $M0$ where d represents the depth of the generated matrix. The procedure of generating the permuted matrices follows a depth search first (DSF) recursicve approach where the stopping condition (leaf matrices) conform to the following form:

(1) M contains only $0's$ and $1's$.

(2) There is exactly one 1 in each row.

(3) Not more than one 1 in each colum.

An demonstration of how the permutation matrices are generated is demonstrated in figure 6.

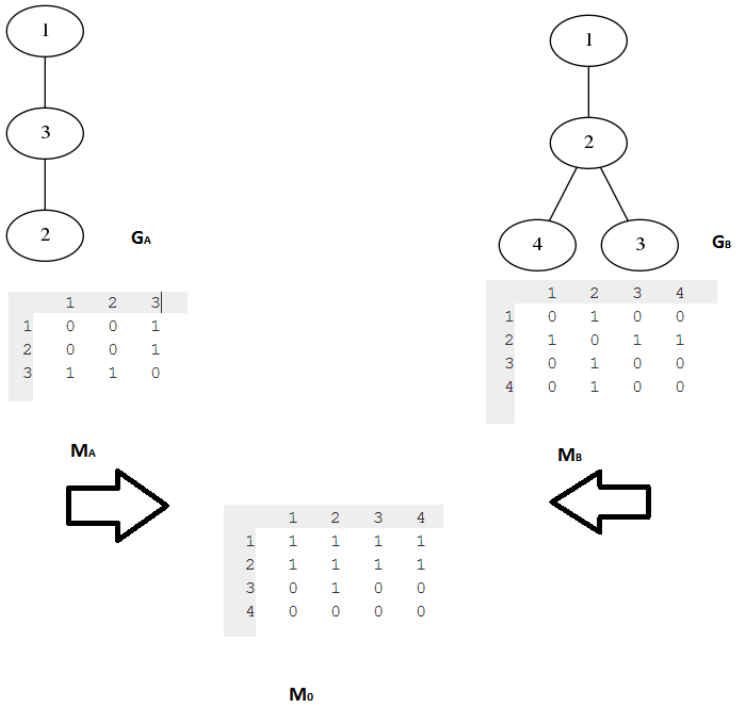


Figure 6: Demonstation of how a permutation matrix is generated from two graphs

Once all the permutation matrices have been generated, each one of the matrices is matched with a graph C, that is obtained from the dot product of the permuted matrix and the graph G_A. The formula for calculating graph C is follows: $C{=}M_n(M_n \ .\ G_A)T$, where G_A = input graph and M_n = permutated matrix M_n in M^d, obtained from the starting matrix $M0$ $(M_n \ .\ G_A)T$ = the transpose of the dot product of the permutated matrix M_n and the graph G_A. If there is a single instance of the matrix C, that is calculated using some permutated matrix M_n obtained from the starting matrix $M0$, that is equal to matrix G_B, then G_B is isomorphic to G_A. Thus G_B is isomorphic to G_A $if f G_B$

$$ij = 1 \rightarrow Cij = 1 \forall i, j \tag{5}$$

If non of the generated permutated matrices can statisfy this condition, then G_B is not isomorphic to G_B.

2.3 VF2 Algorithm

2.3.1 Algorithm

The VF2 algorithm was introduced by L.P.Cordella, P.Foggiaa, C.Sansone and M.Vento [11]. The algorithm is suitable for graph matching and isomorphic determination, including subgraph isomorphic determination on large graphs, this is attributed to the Data structures that the algorithm uses and the manner in which they are used [11], this feature is discussed later in the paper.

The algorithm performs the matching process by attempting to find a mapping M, of vertices in graph G_A which correspond to vertices in graph G_B. The mapping is then used to determine if the two graphs are completely syntactically similar(isomorphic), partially syntactically similar or have no structural similarities at all.

2.3.1.1 Matching and Mapping definition

A mapping M is defined as a set of pairs (n, m), where n is a vertice from G_A and m a vertice from G_B, thus $n \subseteq G_A$ and $m \subseteq G_B$.

The isomorphism determining properties of the mapping are defined as follows, a mapping $M \subset N_A * N_B$ is isomorphic $iff M$ is a bijection, that preserves the branching structure of G_A and G_B, where N_A is a set of vertices from G_A and N_B a set of vertices from G_B.

The mapping $M \subset N_A * N_B$ is subgraph isomorphic $iff M$ is isomorphic to G_A and a subgraph of G_B.

2.3.1.2 Mapping Procedure

The mapping M comprises of state based partial solution morphisms $M(s)$ for each state s.The process of finding the mapping M that is described above uses State Space Representations (SSR) [12]. The partial solution morphisms $M(s)$ selects two subgraphs from G_A and G_B, namely $G_A(s)$ and $G_B(s)$ respectively. The subgraphs comprises of only vertices that are present in the partial solution $M(s)$ for the state s as well as the edges joining them together.

The algorithm starts with an initial state $s0$ that has no mapping between the two graphs, thus $M(s0) = NULL$. The algorithm then computes a set of candidate pairs $P(s)$. Each candidate p in the set is checked against the feasibility function that is discussed in the following chapter, if p is successful then it is added to the state s. And the successor s' is computed using a combination of the predecessor state and the candidate p, thus:

$$s' = s \cup p \tag{6}$$

The process of generating successor states is a recursive procedure that makes use of the depth first traversal for graphs. When a path has been exhausted and a solution has not yet been found, the algorithm uses backtracking to explore the alternative paths [11,13].

2.3.1.3 Definition of the set $P(s)$ and of the feasibility function $F(s, n, m)$

The VF2 algorithm generates the states with close consideration that only some of the states are consistant with the desired morphisms [12]. The algorithm avoids inconsistant states by making use of a set of rules in it's state generation procedure, thus ensuring that only consistant state are generated, these rules are refered to as feasibility rules.

The algorithm uses a function called a feasibility function to test that an additon of a pair (n, m) to a state will be consistant. If the addtion of the pair passes all the feasibility rules, the algorithm will return a true value, if not, a false value indicating that the procedure results in an inconsistant successor state s', and thus that state s' will not be explored by the algorithm.

A further filter can be applied in the consistent states to rule out those states whose successor states will be inconsistant, this apporoach is employed by adding a additional rules called $k-look-ahead$ rules [12]. They check to see if the current state s will have a consistant successor state after k steps, i.e. they check to see if the states from s to s^k are consistant with the desired morphisms.

2.3.1.4 Condidate Pairs

The candidate pairs are obtained by considering the vertices that are connect to $G_A(s)$ and $G_B(s)$, the sub-graphs of G_A and G_B in the state s. The vertexs are used to form the pairs (n, m) as defined above. In order explain how the pairs are formed, we must first introduce the following definitions: Let:

(1) $T_Ain(s)$ be the set of vertexs that are not yet in the partial mapping $M(s)$ and are the origin of the edges from graph G_A

(2) $T_Bin(s)$ be the set of vertexs that are not yet in the partial mapping $M(s)$ and are the origin of the edges from graph G_B

(3) $T_Aout(s)$ be the set of vertexs that are not yet in the partial mapping $M(s)$ and are the destination of the edges from graph G_A

(4) $T_Bout(s)$ be the set of vertexs that are not yet in the partial mapping $M(s)$ and are the destination of the edges from graph G_B

The pair (n, m) is made by vertex n from $T_Aout(s)$ and m from $T_Bout(s)$. If the any of the sets is empty, then we consider the vertex n from $T_Ain(s)$ and m from $T_Bin(s)$. In the case where that graphs are not connected, the pairs will be made by all the vertex not yet contained in either $G_A(s)$ and $G_B(s)$. These pairs form the entries in the set $P(s)$ for that respective state s.

2.3.1.5 The feasibility rules

The feasibility rules that are used to ensure that the states that are evaluated play a role in improving the performance, by preventing inconsistent states from being explored and thus optimizing the execution of the algorithm. There are five general feasibility rules defined as R_{pred},R_{succ},R_{in},R_{out} and R_{new} respectively.

The feasibility functions check for two main things, firstly they check the consistency of the partial solution in the successor state s', namely $M(s')$. Rules R_{pred} and R_{succ} are the rules used for this checking.
The remaining rules are used for pruning the search space for different levels of look ahead. The R_{in} and R_{out} are used to look ahead one level and determine which of those successor states are consistent, and R_{new} is used for the same purpose, but for a look ahead level of two. The definition for each rule is as follows:

$$\begin{aligned} &Rpred(s,n,m) \Longleftrightarrow \\ &\quad (\forall n' \in M1(s) \cap Pred(GB,n) \exists m' \in Pred(GA,m) | (n',m') \exists M(s) \wedge \\ &\quad\quad \forall m' \in M2(s) \cap Pred(GA,m) \exists n' \in Pred(GB,n) | (n',m') \exists M(s)), \end{aligned} \tag{7}$$

$$\begin{aligned} &Rsucc(s,n,m) \Longleftrightarrow \\ &\quad (\forall n' \in M1(s) \cap Succ(GB,n) \exists m' \in Succ(GA,m) | (n',m') \exists M(s) \wedge \\ &\quad\quad \forall m' \in M2(s) \cap Succ(GA,m) \exists n' \in Succ(GB,n) | (n',m') \exists M(s)), \end{aligned} \tag{8}$$

$$\begin{aligned} &Rin(s,n,m) \Longleftrightarrow \\ &\quad (Card(Succ(GB,n) \cap TBin(s)) \geq Card(Succ(GA,m) \cap TAin(s))) \cap \\ &\quad\quad (Card(Pred(GB,n) \cap TBin(s)) \geq Card(Pred(GA,m) \cap TAin(s))) \end{aligned} \tag{9}$$

$$\begin{aligned} &Rout(s,n,m) \Longleftrightarrow \\ &\quad (Card(Succ(GB,n) \cap TBout(s)) \geq Card(Succ(GA,m) \cap TAout(s))) \cap \\ &\quad\quad (Card(Pred(GB,n) \cap TBout(s)) \geq Card(Pred(GA,m) \cap TAout(s))) \end{aligned} \tag{10}$$

$$\begin{aligned} &Rnew(s,n,m) \Longleftrightarrow \\ &\quad Card(\tilde{NA}(s) \cap Pred(GB,n)) \geq Card(\tilde{NB}(s) \cap Pred(GA,n)) \wedge \\ &\quad\quad Card(\tilde{NA}(s) \cap Succ(GB,n)) \geq Card(\tilde{NB}(s) \cap Succ(GA,n)) \end{aligned} \tag{11}$$

In order for a state to be considered consistent, it must pass a combination of all of the five rules, namely:

$$Fsyn(s,n,m) = Rpred \wedge Rsucc \wedge Rin \wedge Rout \wedge Rnew \tag{12}$$

where $F_{\text{syn}}(s,n,m)$ the feasibility function that is envoked upon the state s.

2.4 Conclusion

The Ullman and VF2 algorithm that are discussed above both follow a similar approach in attempting to perform graph mathing. They both construct a tree from the adjacency representation of their input graphs and use the depth first tree traversal techniques to evaluate the graphs, this is done every effectively by both the algorithms.

Though both algorithms are effective in their own respective regards, they are optimized rather differently and thus differe in their degree of complexity. The VF2 algorithm optimizes its execution by performing a look-ahead operation of two states from its current states in an attempt to ignore paths that will result in inconsistant states.

The Ullman algorithm on the other hand optimizes its execution by not computing all possible sub-graphs of some graph G, but reduces the computated matrices by initially computing a matrice $M0$ from the input graphs and then ensuring that all the matrices that are computed from $M0$ are tested so as to prevent the algorithm from exploring a branch that will not result in a graph or sub-graph isomorphism

References

[1] J.R. Ullmann, *"An Algorithm for Subgraph Isomorphism, Journal of the Association for Computing Machinery"*, vol. 23, pp. 31-42, 1976.

[2] H. Ehrig, *"Introduction to Graph Grammars with Applications to Semantic Networks Computers"*, Math. Appl. 23, pp. 557-572, 1992.

[3] D.H. Rouvray, A.T. Balaban, *"Chemical Applications of Graph Theory in Applications of Graph Theory Academic Press"*, New York, pp.177-221, 1979.

[4] S. Melnik, H. Garcia-Molina and E. Rahm, *"Similarity flooding: a versatile graph matching algorithm and its application to schema matching"*, in Proceedings 18th ICDE, San Jose CA, Feb 2002.

[5] L. Cordella, P. Foggia, C. Sansone and M. Vento, *A (sub)graph isomorphism algorithm for matching large graphs IEEE Transactions on Pattern Analysis and Machine Intelligence"*, vol. 26, no. 10, pp. 13671372, Oct. 2004.

[6] *VF2 Algorithm(2015)*, Available at: `https://networkx.github.io/documentation/latest/reference/algorithms.isomorphism.vf2.html`, (Accessed: 10 April 2015).

[7] R. Diestel, *Graph Theory*, Graduate Texts in Mathematics, Vol. 173, Springer Verlag, Berlin, 1991.

[8] A. Drozdek, *Data Structures and Algorithms in Java*, Third Edition,Cengage Learning, pp. 386415, 2013.

[9] Reinhard Diestel, *Graph Theory*, Graduate Texts in Mathematics, Vol. 173, Springer Verlag, Berlin, 1991.

[10] *Graph Definitions*, Available at: `http://web.eecs.utk.edu/~leparker/Courses/CS302-Fall06/Notes/graph-defs-rev.html`, (Accessed: 10 June 2015).

[11] N.J. Nilsson, *Principles of Artificial Intelligence*, Springer-Verlag, 1982.

[12] M. Kubale and B. Jackowski, *A Generalization Implicit Enumeration Algorithm for Graph Coloring*,Communiations of the ACM, 28(4):412-418, 1985.