# COORDINATED SCIENCE LABORATORY

College of Engineering

# Substructure Discovery in SUBDUE

# Lawrence B. Holder

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Approved for Public Release. Distribution Unlimited.

| 1a. REPORT SECURITY CLASSIFICATION       1b. RESTRICTIVE MARKINGS         Unclassified       None         2a. SECURITY CLASSIFICATION AUTHORITY       3. DISTRIBUTION/AVAILABILITY OF REPORT   |   |  |   | REPORT DOCU  | MENTATION  | PAGE  |  |   |
|--|---|--|---|--|--|---|--|---|
| Unclassified       None         None       None         SecURITY CLASSIFICATION AUTHORITY       1. Distribution values         B. DECLASSIFICATION / DOWNGRADING SCHEDULE       1. Distribution values         APPCORMENT ORGANIZATION REPORT NUMBER(S)       1. DILU-ENC-88-2220         UTLU-ENC-88-2220       5. MONTORING ORGANIZATION REPORT NUMBER(S)         UILU-ENC-88-2220       1. DILU-ENC-88-2220         NAME OF FROMMING ORGANIZATION       60. OFFICE SYMBOL<br>(M' applicable)       7. NAME OF MONTORING ORGANIZATION<br>Iniversity of Illinois         NAME OF FUNDING/SOMORING<br>OGGAMAZATION       60. OFFICE SYMBOL<br>(M' applicable)       7. NAME OF MONTORING ORGANIZATION<br>INSF, ONR, DARPA         NSF, ONR, DARPA       7. NAME OF FUNDING/SOMORING<br>OGGAMAZATION       8. OFFICE SYMBOL<br>(M' applicable)       7. NAME OF FUNDING/SOMORING<br>OGGAMAZATION       8. OFFICE SYMBOL<br>(M' applicable)         NSF, ONR, DARPA       7. DAOPESS (CP, Steer, and ZP Code)       1. DATE of REPORT TO DAY A 22202<br>-2308         NSF, ONR, DARPA       8. DORECT FUNDING NUMBERS       NOO014-87-K-0874         NSO C. Street, Washington D.C. 20552       10. DURCE OF FUNDING NUMBERS       NOO014-87-K-0874         1. THE ORCHORT       13. TIME COVERED       10. DURCE OF FUNDING NUMBERS       NOCET         2. PRECOMAL AUTHORS)       Holder, Lawrence B.       1.       S.         3. ANTECT REPORT       13.   | A. REPORT S   | SECURITY CLAS  | SIFICATION  |  |  |   |  |   |
| a. SECURITY CLASSIFICATION AUTHORITY b. DECLASSIFICATION / AUTLABILITY OF REPORT Approved for public release; distribution unlimited proved for public release; distribution unlimited distribution unlimited distribution unlimited distribution di |   |  |   |  |  |   |  |   |
| b. DECLASSIFICATION / DOWNGRADING SCHEDULE       Approved for public release;<br>distribution unlimited         b. DECLASSIFICATION / DOWNGRADING SCHEDULE       Approved for public release;<br>distribution unlimited         b. TERFORMING ORGANIZATION REPORT NUMBER(S)       S. MONITORING ORGANIZATION REPORT NUMBER(S)         b. UTLU-ENC-88-2220       S. MONITORING ORGANIZATION REPORT NUMBER(S)         b. UTLU-ENC-88-2220       NAME OF FEROTIONING ORGANIZATION REPORT NUMBER(S)         b. OFFICE SYMBOL<br>University of Illinois       N/A         c. ADDRESS (GV, Steek, and 2P Code)<br>Urbana, IL 61801       N/A         a. NAME OF FUNDING SPONSORING<br>ORGANIZATION<br>ORGANIZATION       Bb. OFFICE SYMBOL<br><i>Proposed for Street</i> , Nashington D.C., 20552<br>BOO N. Quincy, Arlington VA 22202         BOO C. Street, Nashington D.C., 20552       NST ISTE-35-11170, NOONI4-82-K-1086         BOD C. Street, Nashington VA 22202       NST ISTE-35-11170, NOONI4-82-K-1086         BOD C. Street, Nashington VA 22202-2308       IS JOUNCE OF FUNDING NUMBERS         BOD C. Street, Nashington VA 22202-2308       IS JOUNCE OF FUNDING NUMBERS         BOD C. Street, Nashington VA 22202       IS JOUNCE OF FUNDING NUMBERS         BOD C. Street, Nashington VA 22202-2308       IS JOUNCE OF FUNDING NUMBERS         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a TYPE OF REPORT TEME SCONTINUE ON TAUE (SCHEME TERMS (Continue on reverse if necessary and identify by block number)         7  | 2a. SECURITY CLASSIFICATION AUTHORITY   |  |   |  |  |   |  |   |
| A DECOMPARIANCE ON CONTROLOGING SHOLL       distribution unlimited            (FEFORMING ORGANIZATION REPORT NUMBER(S)         UTLU-ENC-88-2220        S. MONITORING ORGANIZATION GORGANIZATION         CONTINUE OF CONTINUE  |   |  |   |  | Approved for public release;   |   |  |   |
| UTLU-ENC-68-2220         a. NAME OF PERFORMING ORGANIZATION<br>Coordinated Science Lab       6b OFFICE SYMBOL<br>(# applicable)       7a. NAME OF MONITORING ORGANIZATION<br>Diffusersity of Tillinois         C. ADDRESS (GYD, State, and ZP Code)       N/A       NSF, ONR, DARPA         1101 W. Springfield Avenue<br>Urbana, IL 61801       7b. ADDRESS (GYD, State, and ZP Code)       1800 C. Street, Washington D.C., 20552<br>1000 Wilson Blvd, Arlington VA, 22209-2308         a. MAME OF FUNDING/SPONSORING<br>OSF, ONR, DARPA       8b. OFFICE SYMBOL<br>(# applicable)       9. RECURENT WRTHIMET IDENTIFICATION NUMBER<br>NOT 4-22-R-1038         1800 C. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-COB74       10. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-COB74         1800 C. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-COB74       10. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-COB74         1800 C. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS       10. SOURCE OF FUNDING NUMBERS         1010 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF REPORT (Yest, Month, Day)       11. SPAGE COUNT<br>ACCESSION N         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.       10. SUBDUE       11. DATE (Include Security Security)       12. PAGE COUNT<br>May 1988         2. SUPPLEMENTARY NOTATION       YBBDUE, best-first search,       12. PAGE COUNT<br>May 1988       15. PAGE COUNT<br>12.         3. ABSTRACT (Continue on reverse if necessary and identify by block number)   | 26. DECLASSIFICATION / DOWNGRADING SCHEDULE   |  |   |  |  |   |  |   |
| UILU-ENG-88-2220         a NAME OF PERFORMING ORGANIZATION<br>Coordinated Science Lab<br>University of Illinois       6s OFFICE SYMBOL<br>(if applicable)       7s. NAME OF MONITORING ORGANIZATION<br>NSF, ONR, DARPA         c ADDRESS (GV, Stere, and ZIP Code)       1101 W. Springfield Avenue<br>Urbana, IL 61801       7s. NAME OF MONITORING ORGANIZATION<br>NSF, ONR, DARPA         a NAME OF FUNDING/SPONSORING<br>ORGANIZATION<br>ORGANIZATION<br>NSF, ONR, DARPA       8s. OFFICE SYMBOL<br>(if applicable)       9. PROCLEMENT WITHOUGH THEATION NUMBER<br>NO0014-87-K-0874         1800 C. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS<br>NO0014-87-K-0874       NO0014-87-K-0874         2. ADDRESS (CVD, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS<br>NO0014-87-K-0874       NO0014-87-K-0874         1800 C. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS<br>NO0014-87-K-0874       NOCK UNIT<br>NOCK TASK         19. OULCE, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS       NOCK UNIT<br>NOCK TASK       NOCK UNIT<br>NOCK TASK         10. OULCE, State, and ZIP Code)       10. SOURCE OF REPORT (Year, Month, Day)       11. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         3. BASTRACT (CODES       13. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         5. SUPPLEMENTARY NOTATION       11. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         7       COSATI CODES       11. SUBJECT TERMS (Continue on reverse if necessary and ident  | PERFORMI  | NG ORGANIZAT   | TION REPORT NUMB  | ER(S)  | 5. MONITORING  | ORGANIZATION  | REPORT NUM   | ER(S)   |
| a. NAME OF PERCENNES ORGANIZATION<br>Coordinated Science Lab       60 OFFICE SYMBOL<br>(# applicable)       72. NAME OF MONITORING ORGANIZATION<br>NSF, ONR, DARPA         billow W. Springfield Avenue<br>Urbans, IL 61801       N/A       NSF, ONR, DARPA         a. MARE OF FUNDING/SPONSORING<br>ORGANIZATION       8b. OFFICE SYMBOL<br>(# applicable)       76. ADDRESS (City, State, and ZP Code)<br>1800 C. Street, Washington D.C., 20552<br>1000 Wilson Bivd, Arlington VA. 22209-2308         a. MARE OF FUNDING/SPONSORING<br>ORGANIZATION<br>NSF, ONR, DARPA       8b. OFFICE SYMBOL<br>(# applicable)       9. PROCUEMENT NUMERT DEWIFICATION NUMBER<br>NSF DIRT, State, and ZP Code)         1800 C. Street, Washington D.C. 20552<br>800 N. Quincy, Arlington VA. 22202       10. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-0874         1800 C. Street, Washington VA. 22202<br>1400 Wilson Blvd, Arlington VA. 22202-2308       19. SOURCE OF FUNDING NUMBERS<br>NOOL4-87-R-0874         Substructure Discovery in SUBDUE       19. SOURCE OF REPORT (Year, Month, Day)       15. PAGE COUNT<br>12         2. PERSONAL AUTHORS)       Holder, Lawrence B.       14. DATE OF REPORT (Year, Month, Day)       15. PAGE COUNT<br>12         3. TYPE OF REPORT       13. SUBJECT TEMMS (Continue on reverse if necessary and identify by block number)         7       COSATI CODES       18. SUBJECT TEMMS (Continue on reverse if necessary and identify by block number)         7       COSATI CODES       18. SUBJECT TEMMS (Continue on reverse if necessary and identify by block number)         7       COSATI CODES       1  |   |  |   |  | A Bungal   |   |  |   |
| Coordinated Science Lab       (# applicable)<br>N/A       NSF, ONR, DARPA         Coordinated Science Lab       N/A       NSF, ONR, DARPA         Coordinated Science Lab       N/A       NSF, ONR, DARPA         Coordinated Science Lab       N/A       NSF, ONR, DARPA         Libit W. Springfield Avenue<br>Urbana, IL 61801       N/A       NSF, ONR, DARPA         A MAG OF FUNDING/SPONSORING<br>Coordinated Science Lab       Bb. OFFICE SYMBOL<br>(# applicable)       NSF, ONR, DARPA         A MAG OF LUNDING/SPONSORING<br>Coordinated Science Lab       Bb. OFFICE SYMBOL<br>(# applicable)       NSF ONR, DARPA       NSF ONR, DARPA         A ADRESS (Cive, State, and ZIP Code)       Bb. OFFICE SYMBOL<br>(# applicable)       NSF ONR, DARPA       NSF ONR, DARPA         NSF, ONR, DARPA       Coordinated Science Lab       NSF ONR, DARPA       NSF ONR, DARPA         1800 C., Street, Washington D.C. 20552       NOOL4-87-K-0874       NSE Course Funding Numbers         1800 G., Street, Washington D.C. 20552       NOOL4-87-K-0874       NCCE STIMON         Substructure Discovery in SUBDUE       10. SOURCE OF FUNDING NUMBERS       NOCK UNIT         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.       33. TYPE OF REPORT       Task       NSE Course         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.       SUBPLICE, best-first search,       12         3. ASTPACT (C  | 0220  |  |   |  |  |   |  | •   |
| University of Illinois       N/A       NSF, ONR, DARPA         C ADDRESS (Gry, Stee, and ZP Code)       1101 W. Springfield Avenue       1800 G. Street, Washington D.C., 20552         100 N. Quincy, Arlington, VA 22209-2308       1400 Wilson Bivd, Anigton VA, 22209-2308         A. NAME OF FUNDING/SPONSORING<br>ORGANIZATION       8b. OFFICE SYMBOL<br>(M applicable)       PROCEEMENT WSTRUMENT DEWIFFCATION NUMBER<br>NSF, ONR, DARPA         A. NAME OF FUNDING/SPONSORING<br>ORGANIZATION       8b. OFFICE SYMBOL<br>(M applicable)       PROCEEMENT WSTRUMENT DEWIFFCATION NUMBER<br>NSF, STET=53-11170, NUO014-82-K-1086         A. ADDRESS (Gry, Stree, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS       NOO14-27-K-0874         1800 G. Street, Washington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1800 G. Street, Street, Tington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1800 G. Street, Street, Street, Lawrence B.       10. SOURCE OF FUNDING NUMBERS         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TME COVERED         7.       COSATI CODES         18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         7.       COSATI CODES         19. ADSTRACT (Continue on reverse if necessary and identify by block number)         7.       COSATI CODES         19. ADSTRACT (Continue on reverse if necessary and identify by block number)   |   |  |   |  | 7a. NAME OF N  | IONITORING ORC  | GANIZATION   |   |
| ADDRESS (Gby, State, and ZiP Code)         1101 W. Springfield Avenue         Urbana, IL 61801         NAME OF PUNDING (SPONSORING<br>ORGANIZATION         (P) Source OF FUNDING (NUMBERS<br>PROJECT         (P) SOURCE OF FUNDING (NUMBERS<br>PROJECT         (P) OURCE OF FUNDING (NUMBERS<br>PROJECT         (P) OF RESONAL AUTHOR(S)         (P) OF RESONAL AUTHOR(S)   |   |  |   |  | NSE ONR  | DARPA   |  |   |
| 1101 W. Springfield Avenue       1800 C. Street, Washington D. C. 20552         800 M. Quincy, Arlington, VA 22202       1400 Wilson Blvd, Arlington VA 22209-2308         9, PROCUENCEY INSTUDENTS, ONE, ORK, Stat. and ZP Code       8. OFFICE SYMBOL<br>(# applicable)       9, PROCUENCEY INSTUDENT CASTACT         800 C. Street, Washington D. C. 20552       9, PROCUENCEY INSTUDENT CASTACT       10. SOURCE OF FUNDING NUMBERS         800 C. Street, Washington D. C. 20552       10. SOURCE OF FUNDING NUMBERS       INST. ON A 22209-2308         10. NULL (MACUS Score)       10. SOURCE OF FUNDING NUMBERS       INSC. INS   | of some the second second second  |  |   | M/A  |  |   |  |   |
| 100 N. Optimistic M Weinde       800 N. Quincy, Arilington, VA 22202         1400 Wilson Blvd, Arlington VA, 22209-2308         a. NAME OF FUNDING/SPONSORING<br>ORGANIZATION       8b. OFFICE SYMBOL<br>(# applicable)       9, PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>NSF 15T-85-11170, N00014-82-K-10868         NSF, ONR, DARPA       8b. OFFICE SYMBOL<br>(# applicable)       9, PROCUREMENT INSTRUMENT IDENTIFICATION NUMBERS         1800 N. Quincy, Arilington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1800 N. Quincy, Arilington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1400 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1400 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1400 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         1400 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF REPORT NO.         1400 Wilson Blvd, Arlington VA 22209-2308       10. SOURCE OF REPORT NO.         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3. TYPE OF REPORT       13. TIME COVERED<br>FROM         7       COSATI COOES       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         8. SUPPLEMENTARY NOTATION       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         9. ABSTRACT (Co  |   |  |   |  | 1800 G. Str  | reet, Washi   | ngton D.C.   | . 20552   |
| a. NAME OF FUNDING/SPONSORING<br>ORGANIZATION       Bb. OFFICE SYMBOL<br>(# applicable)       PROCUREMENT INSTRUMENT IDENTIFICATION NUMBERS<br>PROCUREMENT INSTRUMENT IDENTIFICATION NUMBERS<br>NSF, ONR, DARPA         NSF, ONR, DARPA       Bb. OFFICE SYMBOL<br>(# applicable)       9. PEOCUREMENT INSTRUMENT IDENTIFICATION NUMBERS<br>NO0014-87-K-0874         2. ADDRESS (Crine, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS<br>PROGRAM       10. SOURCE OF FUNDING NUMBERS<br>PROGRAM         1800 G. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS         200 N. Quincy, Arlington VA 22209-2308       PROJECT<br>NO.       TASK<br>NO.         1. TITLE (Include Security Classification)       SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED<br>FROM       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT<br>12         5. SUPPLEMENTARY NOTATION       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         7.       COSATI CODES       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         8. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE,<br>best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       The<br>method involves a computationally constrained best-first search (not heuristics:<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure.  |   |  |   |  | 800 N. Quin  | ncy, Arling   | ton, VA 22   | 202   |
| NSF, ORR, DARPA       NO0014-87-K-0874         C ADDRESS (City, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS         1800 G. Street, Washington D.C. 20552       PROGRAM         800 N. Quincy, Arlington VA 22202       PROGRAM         1400 Hilson Blvd, Arlington VA 22209-2308       PROGRAM         1. Thte (include Security Classification)       SUBstructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED         Technical       To         14. DATE OF REPORT (Year, Month, Day)       15. PAGE COUNT         12.       SUBPLEMENTARY NOTATION         7.       COSATI CODES       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (continue on reverse if necessary and identify by block number)       SUBDUE system. The method involves a computationally constrained best-first search,         9. ABSTRACT (continue on reverse if necessary and identify by block number)       Substructure selection, chooses the beset substructures. The process by   | orbana  | , IL 018   | 01  |  | 1400 Wilson  | n Blvd, Arl   | ington VA,   | 22209-2308  |
| NSF, ONR, DARPA       N00014-87-K-0874         K. ADDRESS (Gry, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS         1800 G. Street, Washington D.C. 20552       PROGRAM         800 N. Quincy, Arlington VA 22202       PROGRAM         1400 Wilson Blvd, Arlington VA 22209-2308       PROGRAM         1. TITLE (include Security Classification)       Substructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED         Technical       To         14. DATE OF REPORT May 1988       12         7.       COSATI CODES         18. SUBJECT TERMS (Continue on reverse if mecessary and identify by block number)         SUBDUE,       SUBDUE,         best-first search,         9. ABSTRACT (Continue on reverse if mecessary and identify by block number)         SUBDUE,       SUBDUE,         best-first search,         9. ABSTRACT (Continue on reverse if mecessary and identify by block number)         This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search,         9. ABSTRACT (Continue on reverse if mecessary and identify by block number)         This paper describes the substructure generation and substructure selection. Substructure generation is the process by which new substructures are generate   |   |  | ONSORING  |  | 9. PROCUREMEN  | TINSTRUMENT   | IDENTIFICATION   | NUMBER  |
| C. ADDRESS (City, State, and ZIP Code)       10. SOURCE OF FUNDING NUMBERS         1800 G. Street, Washington D.C. 20552       10. SOURCE OF FUNDING NUMBERS         800 N. Quincy, Arlington VA 22209-2308       10. SOURCE OF FUNDING NUMBERS         10. TITLE (Include Security Classification)       WORK UNIT         Substructure Discovery in SUBDUE       11. TITLE (Include Security Classification)         Substructure Discovery in SUBDUE       12. PRESONAL AUTHON(S)         Holder, Lawrence B.       13. TIME COVERED         3a. TYPE OF REPORT       13. TIME COVERED         Technical       13. TIME COVERED         7.       COSATI CODES         18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         SUBDUE,       best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)         SUBDUE,       best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)         SUBDUE,       best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)         Substructure generation and substructure selection. Substructure generation is the process by which new substructure generation from previously considered substructures. The second process, substructure generation and substructure selection processes are described along with their role in the evaluation of a substructure. After the generati   |   |  |   | (If applicable)  | NSF IST-8  | -111/0, NO  | 0014-82-K-   | 1086  |
| 1800 G. Street, Washington D.C. 2052       PROGRAM NO. PROJECT TASK NO.       TASK NO.         800 N. Quincy, Arlington VA 22202       PROGRAM NO.       PROJECT NO.       TASK NO.         800 N. Quincy, Arlington VA 22209-2308       PROJECT NO.       NO.       NO.       MO.         900 N. Quincy, Arlington VA 22209-2308       PROJECT NO.       NO.       MO.       MCCESSION NO.         910 Michael Security Classification)       SUBStructure Discovery in SUBDUE       10.       No.       No.       MCCESSION NO.         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.       3a.       Trechnical       To       14.       DATE OF REPORT (Yeer, Month, Dey)       15.       PAGE COUNT         7.       COSATI CODES       18.       SUBJUE, best-first search,       9.       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,       9.         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search, substructures, condition is the process. substructure generation and substructure selection. Substructure generation is the process. substructure generation and substructure generation is the second process. substructure are generated from previously considered substructures. The second process. substructure are generated from previo   |   |  | 1 710 6- 1-1  | 1  |  |   |  |   |
| 800 N. Quincy, Arlington VA 22202       ELEMENT NO. NO. NO. NO. CERSSION NO.         1400 Wilson Blvd, Arlington VA 22209-2308       NO. NO. NO. CERSSION NO.         1400 Wilson Blvd, Arlington VA 22209-2308       NO. NO. NO. CERSSION NO.         Substructure Discovery in SUBDUE       Substructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED         Technical       TO         7.       COSATI CODES         FIELD       GROUP         SUBJUE,       Dest-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)         This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described along with their role in the evaluation of a substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.   |   |  |   | 00550  |  |   |  |   |
| 1400 Wilson Bivd, Arlington VA 22209-2308         1. TITLE (Include Security Classification)         Substructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S) Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED FROM         Technical       1b. TIME COVERED FROM         7.       COSATI CODES         FIELD       Image: Substructure on reverse if necessary and identify by block number)         SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)         This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings. compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described along with their role in the discovery of substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.         20. DISTRIBUTION/AVALABILITY OF ABSTRACT       21. ABSTRACT SECURITY CLASSIFICATION         21. MOSTRIBUTION/AVALABILITY OF ABSTRACT       21. ABSTRACT SECURITY CLASSIFICATION   |   |  |   |  |  |   |  |   |
| 1. TITLE (Include Security Classification)         Substructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT         Technical       13b. TIME COVERED       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT         12       12       12         6. SUPPLEMENTARY NOTATION       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         7.       COSATI CODES       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure concepts.   |   |  |   |  | 1  | A STATES  |  | The show we   |
| Substructure Discovery in SUBDUE         2. PERSONAL AUTHOR(S)       Holder, Lawrence B.         3a. TYPE OF REPORT       13b. TIME COVERED       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT         Technical       FROM       TO       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT         3b. SUPPLEMENTARY NOTATION       TO       14. DATE OF REPORT (Yeer, Month, Day)       15. PAGE COUNT         7.       COSATI CODES       18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE, best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       Substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure according best-first search guided by four heuristics: State described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure concepts.  |   |  |   |  | -  | a state the second  |  |   |
| FIELD       GROUP       SUB-GROUP       SUBDUE,<br>best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       SUBDUE,<br>best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       This paper describes the substructure discovery method used in the SUBDUE system. The<br>method involves a computationally constrained best-first search guided by four heuristics:<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure selection. Substructure generation is the<br>process by which new substructures are generated from previously considered substructures. The<br>second process, substructure selection, chooses the best substructure among alternative<br>substructures according to the four heuristics. Each of the four heuristics are described along with<br>their role in the evaluation of a substructure. After the generation and selection processes are<br>described, the substructure discovery algorithm is presented. Two examples demonstrate<br>SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.         0. DISTRIBUTION/AVAILABILITY OF ABSTRACT       21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified  | 2. PERSONA<br>3a. TYPE OF<br>Techni   | REPORT   | Holder, Law<br>13b. TIME C<br>FROM  | rence B.   | 14. DATE OF REPO<br>May 198  | DRT (Year, Month  |  |   |
| FIELD       GROUP       SUB-GROUP       SUBDUE,<br>best-first search,         9. ABSTRACT (Continue on reverse if necessary and identify by block number)       This paper describes the substructure discovery method used in the SUBDUE system. The<br>method involves a computationally constrained best-first search guided by four heuristics:<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure selection. Substructure generation is the<br>process by which new substructures are generated from previously considered substructures. The<br>second process, substructure selection, chooses the best substructure among alternative<br>substructures according to the four heuristics. Each of the four heuristics are described along with<br>their role in the evaluation of a substructure. After the generation and selection processes are<br>described, the substructure discovery algorithm is presented. Two examples demonstrate<br>SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.         20. DISTRIBUTION/AVAILABILITY OF ABSTRACT       IDTIC USERS       21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified   | 2. PERSONA<br>3a. TYPE OF<br>Techni   | REPORT   | Holder, Law<br>13b. TIME C<br>FROM  | rence B.   | 14. DATE OF REPO<br>May 198  | DRT (Year, Month<br>8   |  |   |
| SUBDUE,<br>best-first search,<br>9. ABSTRACT (Continue on reverse if necessary and identify by block number)<br>This paper describes the substructure discovery method used in the SUBDUE system. The<br>method involves a computationally constrained best-first search guided by four heuristics:<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure selection. Substructure generation is the<br>process by which new substructures are generated from previously considered substructures. The<br>second process, substructure selection, chooses the best substructure among alternative<br>substructures according to the four heuristics. Each of the four heuristics are described along with<br>their role in the evaluation of a substructure. After the generation and selection processes are<br>described, the substructure discovery algorithm is presented. Two examples demonstrate<br>SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.<br>20. DISTRIBUTION/AVAILABILITY OF ABSTRACT INT CLASSIFICATION<br>MUNCLASSIFIED/UNLIMITED SAME AS RPT. OTIC USERS   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEM  | REPORT   | Holder, Law<br>13b. TIME C<br>FROM  | rence B.<br>OVEREDTO   | May 198  | 38  | 1  | 2   |
| <ul> <li>P. ABSTRACT (Continue on reverse if necessary and identify by block number)</li> <li>This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure concepts.</li> <li>DOISTRIBUTION/AVAILABILITY OF ABSTRACT</li> <li>ABSTRACT SECURITY CLASSIFICATION Unclassified</li> </ul>   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEM<br>7.  | E AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES   | rence B.<br>OVEREDTO   | May 198  | 38  | 1  | 2   |
| This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure concepts.   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.   | E AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES   | TO   | May 198  | 38  | 1  | 2   |
| This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure concepts.   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.   | E AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES   | rence B.<br>OVERED<br>TO<br>18. SUBJECT TERMS (<br>SUBDUE,   | May 198  | 38  | 1  | 2   |
| method involves a computationally constrained best-first search guided by four heuristics.<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure selection. Substructure generation is the<br>process by which new substructures are generated from previously considered substructures. The<br>second process, substructure selection, chooses the best substructure among alternative<br>substructures according to the four heuristics. Each of the four heuristics are described along with<br>their role in the evaluation of a substructure. After the generation and selection processes are<br>described, the substructure discovery algorithm is presented. Two examples demonstrate<br>SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.<br>21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD  | COSATI   | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP  | rence B.<br>OVERED<br>TO<br>18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs  | May 198<br>Continue on reven<br>t search,  | 38  | 1  | 2   |
| method involves a computationally constrained best-first search guided by four heuristics.<br>cognitive savings, compactness, connectivity and coverage. The two main processes contained in<br>this method are substructure generation and substructure selection. Substructure generation is the<br>process by which new substructures are generated from previously considered substructures. The<br>second process, substructure selection, chooses the best substructure among alternative<br>substructures according to the four heuristics. Each of the four heuristics are described along with<br>their role in the evaluation of a substructure. After the generation and selection processes are<br>described, the substructure discovery algorithm is presented. Two examples demonstrate<br>SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.<br>21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD  | COSATI   | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP  | rence B.<br>OVERED<br>TO<br>18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs  | May 198<br>Continue on reven<br>t search,  | 38  | 1  | 2   |
| <ul> <li>cognitive savings. compactness. connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.</li> <li>O. DISTRIBUTION / AVAILABILITY OF ABSTRACT</li> <li>21. ABSTRACT SECURITY CLASSIFICATION Unclassified</li> </ul>  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD  | COSATI<br>GROUP  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary  | TO<br>18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of  | May 198<br>Continue on reven<br>t search,<br>number)   | ie if necessary an  | nd identify by i   | 2<br>block number)  |
| this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.  | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | COSATI<br>GROUP<br>T (Continue on  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>aper describes t  | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block in  | May 198<br>Continue on reven<br>t search,<br>number)<br>covery method  | used in the   | nd identify by a   | 2<br>block number)<br>stem. The   |
| process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.          0. DISTRIBUTION/AVAILABILITY OF ABSTRACT       21. ABSTRACT SECURITY CLASSIFICATION Unclassified   | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>payings compacting  | TO<br>TO<br>18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block is<br>tationally constraint<br>ess. connectivity and  | May 198<br>Continue on reven<br>t search,<br>number)<br>covery method<br>ned best-first<br>ad coverage. T  | used in the search guiden   | 1<br>nd identify by 1<br>SUBDUE sys<br>d by four<br>processes co   | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in  |
| second process. substructure selection. chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.  | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tion are substructure   | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>the substructure disc<br>tationally constrain<br>ess, connectivity and<br>e generation and sul  | May 198<br>Continue on revers<br>t search,<br>number)<br>covery method<br>hed best-first<br>id coverage. T<br>pstructure selec   | used in the search guided<br>he two main  | 1<br>nd identify by 1<br>SUBDUE sys<br>d by four<br>processes co<br>acture genera  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the   |
| <ul> <li>substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.</li> <li>O. DISTRIBUTION / AVAILABILITY OF ABSTRACT</li> <li>QUNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS</li> <li>21. ABSTRACT SECURITY CLASSIFICATION Unclassified</li> </ul>  | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method  | Holder, Law<br>13b. TIME C<br>FROM  | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>tationally constrain<br>ess. connectivity and<br>e generation and sul<br>ructures are general   | May 198<br>Continue on revers<br>t search,<br>number)<br>covery method<br>ned best-first<br>ad coverage. T<br>postructure select<br>ted from previo  | used in the search guided<br>he two main<br>ction. Substru-   | 1<br>SUBDUE system<br>d by four<br>processes concurrence<br>red substruction   | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The   |
| <ul> <li>their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.</li> <li>O. DISTRIBUTION / AVAILABILITY OF ABSTRACT</li> <li>QUNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS</li> <li>21. ABSTRACT SECURITY CLASSIFICATION Unclassified</li> </ul>   | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>.cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by   | Holder, Law<br>13b. TIME C<br>FROM  | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ure selection, cho  | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>ned best-first<br>ad coverage. T<br>postructure select<br>ted from previo<br>oses the best  | used in the search guided<br>he two main<br>ction. Substructure<br>substructure   | 1<br>SUBDUE system<br>d by four<br>processes control<br>true generation<br>red substruction<br>re among at   | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative  |
| <ul> <li>described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.</li> <li>O. DISTRIBUTION / AVAILABILITY OF ABSTRACT</li> <li>21. ABSTRACT SECURITY CLASSIFICATION Unclassified</li> </ul>   | 2. PERSONA<br>Ba. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro   | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tivings, compacting<br>are substructure<br>which new substructure<br>which new substructure<br>to cess, substructure<br>to | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block i<br>tationally constrain<br>ess, connectivity an<br>e generation and sul<br>ructures are generation<br>ure selection, cho<br>he four heuristics.   | May 198<br>Continue on reverse<br>t search,<br>humber)<br>covery method<br>hed best-first<br>d coverage. T<br>ostructure select<br>ted from previo<br>oses the best<br>Each of the fou   | used in the search guided<br>he two main<br>ction. Substruc-<br>busly consider<br>tously consider<br>tously consider<br>tously consider<br>tously consider<br>tously consider                   | 1<br>nd identify by 1<br>SUBDUE sys<br>d by four<br>processes co<br>acture genera<br>red substruct<br>re among<br>re described   | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>along with  |
| SUBDUE's ability to discover substructure and the advantages to be gained by other learning<br>systems from the discovery of substructure concepts.<br>0. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>UNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS<br>Unclassified  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tivings, compacting<br>are substructure<br>which new substructure<br>which new substructure<br>to cess, substructure<br>to | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block i<br>tationally constrain<br>ess, connectivity an<br>e generation and sub<br>ructures are generation<br>ure selection, cho<br>he four heuristics.   | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>d coverage. T<br>ostructure select<br>ted from previo<br>oses the best<br>Each of the fou<br>. After the ge   | used in the search guided<br>he two main<br>ction. Substru-<br>busly consider<br>to substructure<br>in heuristics an<br>neration and  | 1<br>and identify by a<br>SUBDUE syst<br>d by four<br>processes con-<br>acture general<br>red substruct<br>re among a<br>re described<br>selection pro-  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>along with<br>occesses are                              |
| Systems from the discovery of substructure concepts.<br>0. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>UNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS<br>Unclassified   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described                              | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tivings, compacting<br>are substructure<br>which new substructure<br>which new substructure<br>the evaluation<br>the substructure   | 18. SUBJECT TERMS (<br>SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>the substructure disc<br>tationally constrain<br>ess, connectivity and<br>e generation and sub-<br>ructures are generation<br>ure selection, cho<br>he four heuristics.<br>and a substructure<br>the discovery algor | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>id coverage. T<br>ostructure select<br>ted from prevision<br>oses the best<br>Each of the four.<br>After the gest<br>ithm is prese  | used in the search guided<br>he two main<br>stion. Substru-<br>busly consider<br>substructure<br>in heuristics an<br>neration and<br>nted. Two  | 1<br>SUBDUE syst<br>d by four<br>processes co<br>acture generative<br>red substructive<br>re among a<br>red substructive<br>re described<br>selection pro-<br>examples data  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>alternative<br>along with<br>occesses are<br>emonstrate |
| UNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS Unclassified  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>.cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described,<br>SUBDUE's                | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tion are substructure<br>which new substructure<br>which new substructure<br>the substructure<br>the substructure<br>ability to disco   | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ure selection, cho<br>he four heuristics.<br>a of a substructure<br>re discovery algority<br>ver substructure and                   | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>id coverage. T<br>bostructure select<br>ted from previo<br>oses the best<br>Each of the fou<br>. After the ge<br>ithm is prese<br>hed the advanta                                       | used in the search guided<br>he two main<br>stion. Substru-<br>busly consider<br>substructure<br>in heuristics an<br>neration and<br>nted. Two  | 1<br>SUBDUE syst<br>d by four<br>processes co<br>acture generative<br>red substructive<br>re among a<br>red substructive<br>re described<br>selection pro-<br>examples data  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>alternative<br>along with<br>occesses are<br>emonstrate |
| MUNCLASSIFIED/UNLIMITED SAME AS RPT. DTIC USERS Unclassified   | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>.cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described,<br>SUBDUE's                | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tion are substructure<br>which new substructure<br>which new substructure<br>the substructure<br>the substructure<br>ability to disco   | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ure selection, cho<br>he four heuristics.<br>a of a substructure<br>re discovery algority<br>ver substructure and                   | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>id coverage. T<br>bostructure select<br>ted from previo<br>oses the best<br>Each of the fou<br>. After the ge<br>ithm is prese<br>hed the advanta                                       | used in the search guided<br>he two main<br>stion. Substru-<br>busly consider<br>substructure<br>in heuristics an<br>neration and<br>nted. Two  | 1<br>SUBDUE syst<br>d by four<br>processes co<br>acture generative<br>red substructive<br>re among a<br>red substructive<br>re described<br>selection pro-<br>examples data  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>alternative<br>along with<br>occesses are<br>emonstrate |
| Clothe OJERS UNCLASSIFIED  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>5. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT   | L AUTHOR(S)<br>REPORT<br>.cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described,<br>SUBDUE's<br>systems fro | Holder, Law<br>13b. TIME C<br>FROM  | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ure selection, cho<br>he four heuristics.<br>a of a substructure<br>re discovery algority<br>ver substructure and                   | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>id coverage. T<br>bostructure select<br>ted from previo<br>oses the best<br>Each of the fou<br>. After the ge<br>ithm is prese<br>hed the advanta                                       | used in the search guided<br>he two main<br>stion. Substru-<br>busly consider<br>substructure<br>in heuristics an<br>neration and<br>nted. Two  | 1<br>SUBDUE syst<br>d by four<br>processes co<br>acture generative<br>red substructive<br>re among a<br>red substructive<br>re described<br>selection pro-<br>examples data  | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>alternative<br>along with<br>occesses are<br>emonstrate |
| 24. MAINE OF RESPONSIBLE INDIVIDUAL 22b. TELEPHONE (Include Area Code) 22c. OFFICE SYMBOL  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT<br>9. ABSTRACT                            | COSATI<br>COSATI<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described,<br>SUBDUE's<br>systems fro              | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tion are substructure<br>which new substructure<br>which new substructure<br>the substructure<br>ability to disco-<br>om the discovery<br>HUITY OF ABSTRACT   | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>the substructure disc<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ructures are generation<br>in of a substructure<br>of a substructure and<br>of substructure con            | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>d coverage. T<br>ostructure select<br>ted from prevision<br>oses the best<br>Each of the four.<br>After the ge<br>ithm is prese<br>hed the advanta<br>cepts.                            | used in the search guided<br>he two main<br>ction. Substru-<br>busly consider<br>substructure<br>in heuristics an<br>neration and<br>nted. Two<br>ages to be gain                               | 1<br>and identify by a<br>SUBDUE syst<br>d by four<br>processes con-<br>acture general<br>red substruct<br>re among a<br>red substruct<br>re among a<br>re described<br>selection pro-<br>examples da<br>ined by other | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>alternative<br>along with<br>occesses are<br>emonstrate |
|  | 2. PERSONA<br>3a. TYPE OF<br><u>Techni</u><br>6. SUPPLEMI<br>7.<br>FIELD<br>9. ABSTRACT<br>9. ABSTRACT<br>20. DISTRIBU<br>EUNCLAS | L AUTHOR(S)<br>REPORT<br>Cal<br>ENTARY NOTA<br>COSATI<br>GROUP<br>T (Continue on<br>This p<br>method inv<br>cognitive sa<br>this method<br>process by<br>second pro<br>substructur<br>their role i<br>described,<br>SUBDUE's<br>systems fro  | Holder, Law<br>13b. TIME C<br>FROM<br>TION<br>CODES<br>SUB-GROUP<br>reverse if necessary<br>paper describes the<br>volves a compu-<br>tion are substructure<br>which new substructure<br>which new substructure<br>the substructure<br>ability to disco<br>om the discovery<br>HUITY OF ABSTRACT<br>TED SAME AS   | 18. SUBJECT TERMS (<br>SUBDUE,<br>best-firs<br>and identify by block of<br>the substructure disc<br>tationally constrain<br>ess, connectivity and<br>e generation and sub<br>ructures are generation<br>ructures are generation<br>in of a substructure<br>of a substructure and<br>of substructure con            | May 198<br>Continue on reverse<br>t search,<br>number)<br>covery method<br>hed best-first<br>d coverage. T<br>bstructure select<br>ted from previs<br>oses the best<br>Each of the fou<br>. After the ge<br>ithm is prese<br>hd the advanta<br>cepts.<br>21. ABSTRACT SE<br>Unclassi | used in the search guided<br>he two main<br>stion. Substru-<br>busly consider<br>t substructur<br>in heuristics an<br>neration and<br>inted. Two<br>ages to be gain<br>CURITY CLASSIFI-<br>fied | 1<br>and identify by a<br>SUBDUE syst<br>d by four<br>processes con-<br>acture general<br>red substruct<br>re among a<br>re described<br>selection pro-<br>examples da<br>ined by other<br>ICATION                     | 2<br>block number)<br>stem. The<br>heuristics:<br>ntained in<br>tion is the<br>tures. The<br>alternative<br>along with<br>ocesses are<br>emonstrate<br>er learning  |

| UNICLASSIFIE | D |
|--------------|---|
|--------------|---|

# Substructure Discovery in SUBDUE\*

Lawrence B. Holder

Artificial Intelligence Research Group Coordinated Science Laboratory University of Illinois at Urbana-Champaign 1101 West Springfield Avenue Urbana, IL 61801

Telephone: (217) 333-9220 Internet: holder@uicsl.csl.uiuc.edu

March 1988

#### ABSTRACT

This paper describes the substructure discovery method used in the SUBDUE system. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. The two main processes contained in this method are substructure generation and substructure selection. Substructure generation is the process by which new substructures are generated from previously considered substructures. The second process, substructure selection, chooses the best substructure among alternative substructures according to the four heuristics. Each of the four heuristics are described along with their role in the evaluation of a substructure. After the generation and selection processes are described, the substructure discovery algorithm is presented. Two examples demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts.

<sup>&</sup>lt;sup>\*</sup> This research was partially supported by the National Science Foundation under grant NSF IST-85-11170, the Office of Naval Research under grant N00014-82-K-0186, by the Defense Advanced Research Projects Agency under grant N00014-87-K-0874, and by a gift from Texas Instruments, Inc.

#### 1. Introduction

The amount of detailed information available from a real-world environment is overwhelming. Yet, humans have the ability to ignore minute detail and extract information from the environment at a level of detail that is appropriate for the purpose of the observation [Witkin83]. Machine learning systems that operate in such a detailed structural environment must be able to abstract over unnecessary detail in the input and determine which attributes are relevant to the learning task.

Substructure discovery is the process of identifying concepts describing interesting and repetitive "chunks" of structure within structural descriptions of the environment. Once discovered, the substructure concept can be used to simplify the descriptions by replacing all occurrences of the substructure with a single form that represents the newly discovered concept. The discovered substructure concepts allow abstraction over detailed structure in the original descriptions and provide new, relevant attributes for subsequent learning tasks.

This paper describes the substructure discovery method used in the SUBDUE system [Holder88]. The SUBDUE system consists of a substructure discovery module, a substructure specialization module for specializing the substructures discovered by SUBDUE, and an incremental substructure background knowledge module that retains previously discovered substructures for use in subsequent learning tasks. Only the discovery module of SUBDUE is presented in this paper. Section 2 defines *substructure* and related terms. Section 3 discusses the substructure generation process, and Section 4 defines the heuristics used in the substructure selection process. Section 5 outlines SUBDUE's substructure discovery algorithm, and Section 6 illustrates some examples of SUBDUE's performance. Finally, Section 7 summarizes the substructure discovery process in SUBDUE and discusses future work.

#### 2. Substructure

In a graphical sense, a substructure is a collection of nodes and edges comprising a connected subgraph of a larger graph. However, the substructures discovered by SUBDUE represent more than just a syntactic definition of a subgraph. Substructures are concepts. Substructure discovery is concerned with identifying substructures that represent interesting concepts, not just interesting graphical structure. Thus, substructures, or equivalently substructure concepts, should be interpreted as both collections of structurally related objects and as the conjunctive concepts describing them.

An appropriate language for describing substructures is an extension to the first order logic called Variable-valued Logic system 2  $(VL_2)$  [Michalski80], which is a subset of the Annotated Predicate Calculus (APC) [Michalski83a]. Figure 1 illustrates an input example along with the substructure discovered by SUBDUE. Both the input example and the substructure are expressed in the same substructure description language. The expression for the input example shown in

#### Figure 1 is

 $<[SHAPE(T1)=TRIANGLE][SHAPE(T2)=TRIANGLE][SHAPE(T3)=TRIANGLE] \\ [SHAPE(T4)=TRIANGLE][SHAPE(S1)=SQUARE][SHAPE(S2)=SQUARE] \\ [SHAPE(S3)=SQUARE][SHAPE(S4)=SQUARE][SHAPE(R1)=RECTANGLE] \\ [SHAPE(C1)=CIRCLE][COLOR(T1)=RED][COLOR(T2)=RED][COLOR(T3)=BLUE] \\ [COLOR(T4)=BLUE][COLOR(S1)=GREEN][COLOR(S2)=BLUE][COLOR(S3)=BLUE] \\ [COLOR(S4)=RED][ON(T1,S1)=T][ON(S1,R1)=T][ON(C1,R1)=T][ON(R1,T2)=T] \\ [ON(R1,T3)=T][ON(R1,T4)=T][ON(T2,S2)=T][ON(T3,S3)=T][ON(T4,S4)=T]>$ 

If each object of the substructure is assigned a symbolic name as in Figure 1 (e.g., OBJECT-0001, OBJECT-0002), then the expression for the substructure is

<[SHAPE(OBJECT-0001)=TRIANGLE][SHAPE(OBJECT-0002)=SQUARE] [ON(OBJECT-0001,OBJECT-0002)=T]>

A substructure is either a single object or a non-empty set of connected relations. The relations of a substructure are connected if the graph representation of the substructure, where objects are nodes and relations are edges in the graph, is connected. A selector relation consists of the selector relation name, a non-empty set of objects as arguments and the value of the selector relation. Selector relations are henceforth referred to as relations. An object is a primitive element from which relations and, ultimately, substructures are defined.

For the following discussions, some terminology is needed to describe important aspects of substructures as they relate to a given set of input examples. An *occurrence* of a substructure in a set of input examples is a set of objects and relations from the examples that match, graph theoretically, to the graphical representation of the substructure. For example, the occurrences of the substructure in the input example of Figure 1 are

> <[ON(T1,S1)=T][SHAPE(T1)=TRIANGLE][SHAPE(S1)=SQUARE]> <[ON(T2,S2)=T][SHAPE(T2)=TRIANGLE][SHAPE(S2)=SQUARE]> <[ON(T3,S3)=T][SHAPE(T3)=TRIANGLE][SHAPE(S3)=SQUARE]> <[ON(T4,S4)=T][SHAPE(T4)=TRIANGLE][SHAPE(S4)=SQUARE]>

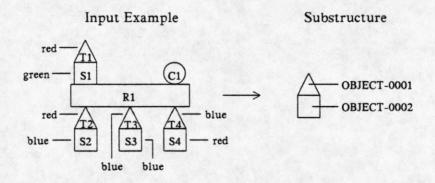


Figure 1. Example Substructure

A neighboring relation of an occurrence of a substructure is a relation in the input example that is not contained in the occurrence, but has at least one object from the occurrence as an argument. For example, the neighboring relations of the first occurrence listed above are [COLOR(T1)=RED], [COLOR(S1)=GREEN] and [ON(S1.R1)=T].

An external connection of an occurrence of a substructure is a neighboring relation of the occurrence that has as an argument at least one object not contained in the occurrence. In other words, an external connection of an occurrence of a substructure is a relation that relates one or more objects in the occurrence to one or more objects not in the occurrence. For the first occurrence listed above, there is only one external connection, [ON(S1,R1)=T].

#### 3. Substructure Generation

An essential function of any substructure discovery system is the generation of alternative substructures. The substructure generation process constructs new substructures from the objects and relations in the input examples. SUBDUE's substructure discovery algorithm employs an approach to substructure generation called minimal expansion. An expansion approach begins with smaller substructures and expands them by appending additional structure from the input examples. Minimal expansion expands the substructures by appending the smallest amount of additional structure. In the context of substructures, this is equivalent to adding one neighboring relation. Thus, minimally expanding a substructure to form a new substructure involves appending one neighboring relation to the substructure. For example, according to the three neighboring relations of the occurrence, <[ON(T1,S1)=T] [SHAPE(T1)=TRIANGLE] [SHAPE(S1)=SQUARE]>, the substructure in Figure 1 would be expanded to generate the following three substructures

> <[SHAPE(OBJECT-0001)=TRIANGLE][SHAPE(OBJECT-0002)=SQUARE] [ON(OBJECT-0001,OBJECT-0002)=T][COLOR(OBJECT-0001)=RED]> <[SHAPE(OBJECT-0001)=TRIANGLE][SHAPE(OBJECT-0002)=SQUARE] [ON(OBJECT-0001,OBJECT-0002)=T][COLOR(OBJECT-0002)=GREEN]> <[SHAPE(OBJECT-0001)=TRIANGLE][SHAPE(OBJECT-0002)=SQUARE] [ON(OBJECT-0001,OBJECT-0002)=T][ON(OBJECT-0002,OBJECT-0003)=T]>

SUBDUE uses an exhaustive minimal expansion technique for generating alternative substructures from a single substructure. The exhaustive version of this technique generates new substructures by considering all possible neighboring relations of the original substructure. To avoid the combinatorical explosion of this process, SUBDUE uses the substructure selection process to select the most promising substructure for expansion.

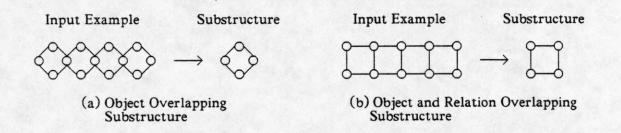
#### 4. Substructure Selection

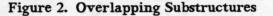
After using the method from the previous section to construct a set of alternative substructures, SUBDUE's substructure discovery algorithm chooses one of these substructures as the best hypothetical substructure. This is the task of substructure selection. The method of selection employs a heuristic evaluation function to order the set of alternative substructures based on their heuristic quality. This section presents the four heuristics used by SUBDUE to evaluate a substructure: cognitive savings, compactness, connectivity and coverage.

The first heuristic, cognitive savings, is the underlying idea behind several utility and data compression heuristics employed in machine learning [Minton87, Whitehall87, Wolff82]. Cognitive savings measures the amount of data compression obtained by applying the substructure to the input examples. In other words, the cognitive savings of a substructure represents the net reduction in complexity after considering both the reduction in complexity of the input examples after replacing each occurrence of the substructure by a single conceptual entity and the gain in complexity associated with the conceptual definition of the new substructure. The reduction in complexity of the input examples can be computed as the number of occurrences of the substructure multiplied by the complexity of the substructure. Thus, the cognitive savings of a substructure, S, for a set of input examples, E, is computed as

cognitive\_savings(S,E) = complexity\_reduction(S,E) - complexity(S) = [number\_of\_occurrences(S,E) \* complexity(S)] - complexity(S) = complexity(S) \* [number\_of\_occurrences(S,E) - 1]

In the above computation of cognitive savings the complexity of the substructure is typically a function of the number of objects, the number of relations, and the arity of the relations in the substructure. However, the number of occurrences of the substructure is more complicated to measure, because occurrences may overlap in the input examples. For instance, Figure 2 shows two input examples along with the substructure found by the discovery process. Here, the circles represent objects and the lines represent relations. At first glance, the number of occurrences of the substructure in Figure 2a may appear to be four; however, the number of non-overlapping occurrences is less than four. Figure 2a illustrates the problem of object overlap, and Figure 2b illustrates the problem of cocurrences of unique occurrences.





In SUBDUE's substructure discovery algorithm, the *complexity*(S) is defined as the size of the substructure, S, where the size is computed as the sum of the number of objects and relations in the substructure. As discussed above, the *number\_of\_occurrences*(S,E) is more complicated to compute, because the occurrences may overlap in the input examples. In view of the overlap problem, simply counting all objects and relations in the overlapping occurrences would incorrectly state the true cognitive savings of the substructure. Therefore, the *complexity\_reduction*(S,E) is redefined to be the number of objects and relations in the occurrences of the substructure, where overlapping objects and relations are counted only once. The number of such objects is referred to as *#unique\_objects*, and the number of such relations is referred to as *#unique\_objects*, the cognitive savings of a substructure, S, with occurrences, OCC, in the set of input examples, E, is computed as

```
cognitive_savings(S,E) = complexity_reduction(S,E) - complexity(S)
```

- = [#unique\_objects(OCC) + #unique\_relations(OCC)] complexity(S)
- = [#unique\_objects(OCC) + #unique\_relations(OCC)] size(S)
- = [#unique\_objects(OCC) + #unique\_relations(OCC)] [#objects(S) + #relations(S)]

As an example of the cognitive savings calculation, consider the input examples and corresponding substructures in Figure 2. If each circle is considered an object and each line a relation, then for each of the two substructures, #objects(S) = 4, #relations(S) = 4, and there are four occurrences of the substructure in the input example. In Figure 2a,  $\#unique_objects(OCC) = 13$  and  $\#unique_relations(OCC) = 16$ ; thus, cognitive\_savings = [13 + 16] - [4 + 4] = 21. In Figure 2b,  $\#unique_objects(OCC) = 10$  and  $\#unique_relations(OCC) = 13$ ; thus, cognitive\_savings = [10 + 13] - [4 + 4] = 15.

The second heuristic, *compactness*, measures the "density" of a substructure. This is not density in the physical sense, but the density based on the number of relations per number of objects in a substructure. The compactness heuristic is a generalization of Wertheimer's *Factor of Closure*, which states that human attention is drawn to closed structures [Wertheimer39]. Graphically, a closed substructure has at least as many relations as objects, whereas a non-closed substructure has fewer relations than objects [Prather76]. Thus, closed substructures have a higher compactness value. Compactness is defined as the ratio of the number of relations in the substructure to the number of objects in the substructure.

 $compactness(S) = \frac{\#relations(S)}{\#objects(S)}$ 

For each of the substructures in Figure 2, #relations(S) = 4 and #objects(S) = 4; thus, compactness = 4/4 = 1.

The third heuristic, connectivity, measures the amount of external connection in the occurrences of the substructure. The connectivity heuristic is a variant of Wertheimer's Factor of Proximity [Wertheimer39], and is related to earlier numerical clustering techniques [Zahn71]. These works demonstrate the human preference for "isolated" substructures, that is, substructures that are minimally related to adjoining structure. Connectivity measures the "isolation" of a substructure by computing the average number of external connections over all the occurrences of the substructure in the input examples. The number of external connections is to be minimized; therefore, the connectivity value is computed as the inverse of the average to arrive at a value that increases as the number of external connections decreases. Thus, the connectivity of a substructure, S, with occurrences, OCC, in the set of input examples, E, is computed as

connectivity(S.E) = 
$$\frac{\left| \sum_{i \in OCC} \left| \text{ external\_connections(i)} \right| \right|^{-1}}{\left| \text{ OCC} \right|}$$

Again consider Figure 2. Each substructure has four occurrences in the input example. For both substructures the two innermost occurrences both have 4 external connections and the two outermost occurrences both have 2 external connections, for a total of 12 external connections. Thus, connectivity =  $(12/4)^{-1}$  = 1/3.

The final heuristic, *coverage*, measures the amount of structure in the input examples described by the substructure. The coverage heuristic is motivated from research in inductive learning and provides that concept descriptions describing more input examples are considered better [Michalski83b]. Coverage is defined as the number of unique objects and relations in the occurrences of the substructure divided by the total number of objects and relations in the input examples. Thus, the coverage of a substructure, S, with occurrences, OCC, in the set of input examples, E, is computed as

$$coverage(S,E) = \frac{\#unique\_objects(OCC) + \#unique\_relations(OCC)}{\#objects(E) + \#relations(E)}$$

For both substructures in Figure 2 the occurrences of the substructure describe every object and relation in the input example; thus, coverage = 1.

Ultimately, the value of a substructure, S, for a set of input examples, E, is computed as the product of the four heuristics.

In this way the compactness, connectivity and coverage heuristics refine the cognitive savings by increasing or decreasing the total value to reflect specific qualities of the substructure. Thus, for the substructure in Figure 2a, value =  $21 \times 1 \times 1/3 \times 1 = 7.0$ ; and for the substructure in Figure 2b, value =  $15 \times 1 \times 1/3 \times 1 = 5.0$ . Applying the heuristic evaluation to the substructure of Figure 1, value =  $15 \times 3/2 \times 1/3 \times 20/37 = 4.054$ .

#### 5. Substructure Discovery Algorithm

Ideally, an algorithm for discovering substructure should converge on the best substructure in terms of the goal of the discovery task. The goal of the substructure discovery algorithm, in general, is to identify the substructure in the input examples that maximizes the capacity for complexity reduction and maximizes the interestingness of the substructure concept. SUBDUE measures both these characteristics with the heuristic evaluation function defined in Section 4. However, the number of possible substructures is exponential in the number of relations within the given input examples. If left unconstrained, the algorithm may eventually consider all possible substructures. SUBDUE imposes a computational limit on the algorithm to constrain the number of substructures considered.

The substructure discovery algorithm used by SUBDUE is a computationally constrained best-first search guided by the substructure generation and selection processes. The algorithm is given one or more input examples and a limit on the amount of computation performed. The algorithm begins by forming the set, S, of alternative substructures. Initially, the set has only one element, the substructure corresponding to a single object, with as many occurrences as there are objects in the input examples. As the algorithm progresses, the discovered substructures are kept in the set, D, which is initially empty.

The next step in the algorithm is a loop that continuously generates new substructures from the substructures in S until either the computational limit is exceeded or the set of alternative substructures, S, is exhausted. The loop begins by selecting the best substructure in S. Here, the value computation of Section 4 is employed to choose the best substructure from the alternatives in S. Once selected, the best substructure is stored in BESTSUB and removed from S. Next, if BESTSUB does not already reside in the set D of discovered substructures, then BESTSUB is added to D. The substructure generation method of Section 3 is then used to construct a set of new substructures by minimally expanding BESTSUB. The newly generated substructures that have not already been considered by the algorithm are added to S, and the loop repeats. When the loop terminates, D contains the set of discovered substructures.

Thus, the substructure discovery algorithm searches for the heuristically best substructure until all possible substructures have been considered or the amount of computation exceeds the given limit. Due to the large number of possible substructures, the algorithm typically exhausts the allotted computation before considering all possible substructures. Therefore, the algorithm may not find the substructure that maximizes the heuristic evaluation function. However, experiments in a variety of domains indicate that the heuristics perform well in guiding the search toward more promising substructures [Holder88].

#### 6. Examples

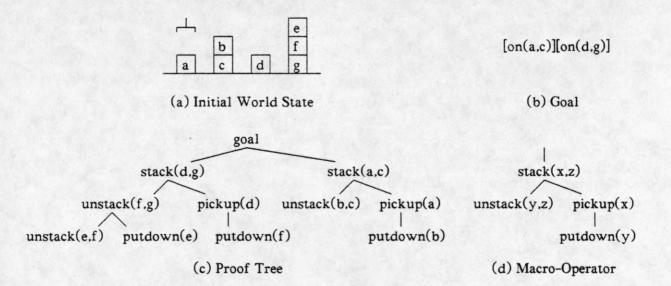
This section presents two examples that demonstrate SUBDUE's ability to discover substructure and the advantages to be gained by other learning systems from the discovery of substructure concepts. Each example is run on a Texas Instruments Explorer using a Common Lisp implementation of the SUBDUE system.

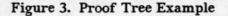
#### 6.1. Example 1

Example 1 illustrates a possible application of the substructure discovery algorithm to the task of discovering macro-operators in plans. The example is drawn from the "blocks world" domain. The operators for this domain are taken from [Nilsson80]: pickup, putdown, stack and unstack.

For this example, suppose the initial world state is as shown in Figure 3a, and the desired goal is in Figure 3b. The proof tree of operators to achieve the goal is shown in Figure 3c. With this proof tree as input, SUBDUE discovers the substructure shown in Figure 3d after considering 19 alternative substructures. The substructure represents a macro-operator for accomplishing a subgoal to stack a block, x, on another block, z, when a block, y, is already on top of block z.

Macro-operators discovered by SUBDUE can be used in several ways. Replacing the occurrences of the macro-operator in the original proof tree by instantiations of the macro-operator can reduce the storage requirements of the schema constructed from the entire proof tree. Retaining the macro-operators discovered within a proof tree would provide sub-schemas in addition to the schemas learned by an explanation-based learning (EBL) system [DeJong86, Mitchell86]. The sub-schemas would increase the amount of operationalized





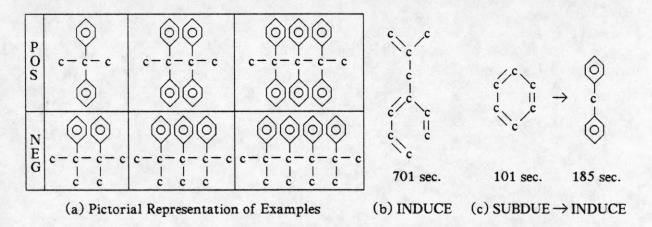
knowledge available to the EBL system for explaining subsequent examples.

#### 6.2. Example 2

Example 2 combines SUBDUE with the INDUCE system [Hoff83] to demonstrate the improvement gained in both processing time and quality of results when the examples contain a large amount of structure. A Common Lisp version of INDUCE was used for this example running on the same Texas Instruments Explorer as the SUBDUE system.

Figure 4a shows a pictorial representation of the three positive and three negative examples given to INDUCE. Each of the symbolic benzene rings in Figure 4a represents the more complex structure in the left side of Figure 4c. The actual input specification for the six examples contains a total of 178 relations of the form [SINGLE-BOND(C1,C2)=T] or [DOUBLE-BOND(C1,C2)=T]. After 701 seconds of processing time, INDUCE produces the concept shown in Figure 4b. Next, all six examples are given to SUBDUE using the same 178 relations. After considering seven alternative substructures for 101 seconds of processing time, SUBDUE discovers the substructure concept of a benzene ring as shown on the left side of Figure 4c. The newly discovered substructure is then used to reduce the complexity of the original examples by replacing each occurrence of the benzene ring with a single relation, i.e., [BENZENE-RING(C1,C2,C3,C4,C5,C6)=T]. Using the reduced set of positive and negative examples, INDUCE produces the concept on the right side of Figure 4c in 185 seconds of processing time. Here, the symbolic benzene rings represent the BENZENE-RING relation, not the complex structural representation used in the original descriptions of the examples.

By abstracting over the structure representing the benzene ring, SUBDUE allows INDUCE to discover the true concept distinguishing the positive and negative examples; namely, benzene rings are paired across one carbon atom in the positive examples, but not in the negative examples. INDUCE represents this





concept in terms of the abstract benzene ring feature provided by SUBDUE. Furthermore, the processing time of SUBDUE and INDUCE combined (286 seconds) represents a speedup of 2.5 over that of INDUCE alone. This example demonstrates how the substructures discovered by SUBDUE can improve the results of other learning systems by abstracting over detailed structure in the input and providing new features.

#### 7. Conclusion

This paper describes the method used by the SUBDUE system to discover substructures in structured examples. The method involves a computationally constrained best-first search guided by four heuristics: cognitive savings, compactness, connectivity and coverage. Alternative substructures are generated by the minimal expansion technique that constructs new substructures by adding minimal structure to previously considered substructures. The two examples demonstrate SUBDUE's ability to find plausible substructures and the possible uses of these substructures by other learning systems.

Earlier work in substructure discovery can be found in Winston's ARCH program [Winston75]. Winston used several domain dependent methods to identify recurring structure in the blocks world examples. Recent work in substructure discovery includes Whitehall's PLAND system for discovering substructure in action sequences [Whitehall87]. Whitehall uses the cognitive savings heuristic along with three levels of background knowledge to discover loops and conditionals in the sequences.

In addition to the substructure discovery module, SUBDUE also contains a substructure specialization module and a substructure background knowledge module. Substructures discovered by SUBDUE are specialized by adding additional structure. Both the original and specialized substructures are stored hierarchically in the background knowledge. The background knowledge may then direct the discovery process towards substructures similar to those already known. Future work and experimentation is necessary to evaluate the improvements gained by using the specialization and background knowledge modules and to incorporate other forms of background knowledge into SUBDUE's substructure discovery process.

#### REFERENCES

#### [DeJong86]

G. F. DeJong and R. J. Mooney, "Explanation-Based Learning: An Alternative View," *Machine Learning 1*, 2 (April 1986), pp. 145-176. (Also appears as Technical Report UILU-ENG-86-2208, AI Research Group, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign.)

#### [Hoff 83]

W. A. Hoff, R. S. Michalski and R. E. Stepp, "INDUCE 3: A Program for Learning Structural Descriptions from Examples," Technical Report UIUCDCS-F-83-904, Department of Computer Science, University of Illinois, Urbana, IL, 1983.

#### [Holder88]

L. B. Holder, "Discovering Substructure in Examples," M.S. Thesis, Department of Computer Science, University of Illinois, Urbana, IL, 1988.

#### [Michalski80]

R. S. Michalski, "Pattern Recognition as Rule-Guided Inductive Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2, 4 (July 1980), pp. 349-361.

# [Michalski83a]

R. S. Michalski, "A Theory and Methodology of Inductive Learning," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, T. M. Mitchell (ed.), Tioga Publishing Company, Palo Alto, CA, 1983, pp. 83-134.

## [Michalski83b]

R. S. Michalski and R. E. Stepp, "Learning from Observation: Conceptual Clustering," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell and T. M. Mitchell (ed.), Tioga Publishing Company, Palo Alto, CA, 1983, pp. 331-363.

#### [Minton87]

S. Minton, J. G. Carbonell, O. Etzioni, C. A. Knoblock and D. R. Kuokka, "Acquiring Effective Search Control Rules: Explanation-Based Learning in the PRODIGY System," *Proceedings of the 1987 International Machine Learning Workshop*, Irvine, CA, June 1987, pp. 122-133.

#### [Mitchell86]

T. M. Mitchell, R. Keller and S. Kedar-Cabelli, "Explanation-Based Generalization: A Unifying View," *Machine Learning 1*, 1 (January 1986), pp. 47-80.

#### [Nilsson80]

N. J. Nilsson, Principles of Artificial Intelligence, Tioga Publishing Company,

#### Palo Alto, CA, 1980.

#### [Prather76]

R. Prather, Discrete Mathematical Structures For Computer Science, Houghton Mifflin Company, New York, NY, 1976.

#### [Wertheimer39]

M. Wertheimer, "Laws of Organization in Perceptual Forms," in A Source Book of Gestalt Psychology, W. D. Ellis (ed.), Harcourt, Brace and Company, New York, NY, 1939.

# [Whitehall87]

B. L. Whitehall, "Substructure Discovery in Executed Action Sequences," M.S. Thesis; Department of Computer Science, University of Illinois, Urbana, IL, 1987. (Also appears as Technical Report UILU-ENG-87-2256)

# [Winston75]

P. H. Winston, "Learning Structural Descriptions from Examples," in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, New York, NY, 1975, pp. 157-210.

### [Witkin83]

A. P. Witkin and J. M. Tenenbaum, "On the Role of Structure in Vision," in *Human and Machine Vision*, J. Beck, B. Hope and A. Rosenfeld (ed.), Academic Press, New York, NY, 1983, pp. 481-543.

# [Wolff82]

J. G. Wolff, "Language Acquisition, Data Compression and Generalization," Language and Communication 2, 1 (1982), pp. 57-89.

#### [Zahn71]

C. T. Zahn, "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters," *IEEE Transactions on Computers C-20*, 1 (January 1971), pp. 68-86.