'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

System for Locating Record Duplication Using Genetic Programming

Mr. Satish Wagh, Prof. Dr. Varsha Patil,
DepartmentofComputerEngineering,
MatoshriCollegeofEngineeringandResearchCenter,Eklahere,Nashik(MS),India, MCERC,Nashik,

***Abstract*— Withrapidadvancementintechnologyenableshighusesofdatabasewhichcausesduplicationofdatabaseman-agement.Thereplicateddatarecordsgeneratemultiplecopiesofdirtydataandcontainimproperspellingandpunctuation,incorrectdataisassociatedwithrecord,incompletedandalsooutdatedorsamedataisduplicatedinthedatabase.forremovingthedirty datavariousdatacleaningsoftwareandapplicationsareused.theproposedgenericprogrammingapproachtorecorddeduplicationhelpsincombiningseveraldifferentpiecesofevidenceextracted fromthe data deduplicationfunction helpsinidentifyingwhethertwoentriesareinrepositoryreplicasornot.Inthis approachforautomatic adaptation of functionaregivenfixedreplicasidentificationboundaryforimprovingaccuracyintermsofnumberofduplicaterecordsfoundversustheactualnumberofduplicaterecords.Thisworkhasbeentestedonthebloodbankdatabasetoduplicatetherecords.Systemusedforlocatingrecorddeduplicationwiththehelpofgenetic programminghelpsincombiningevidenceextractedfromthecontentanddeduplicationfunctionenablesinidentificationweather twoentriespresentin repositoryarereplicasor not. Additionallygeneticprogrammingiscapableforautomaticallyadaptingthesefunctiontoaspecifiedreplicaidentification boundarythisgeneticprogrammingapproachisappliedforvariousdatabasemanagementtofinddeduplicatedrecords.**

***IndexTerms*—Database integration,Evolutionary computingandGeneticalgorithms,Databaseintegration**

I. INTRODUCTION

SEVERAL systemssuch as digital libraries and otherdatabasesystemslikeorganizationdatabasesareaffected bytheduplicates.Thesystemforlocatingrecorddeduplicationusinggeneticprogrammingapproachtofind a deduplica-tionfunctionthat is able to identify whether twoentriesinarepositoryarereplicasornot.Deduplicationisatasko f identifyingthe duplicate data in a repository that refertothesamerealworldentityorobjectandsystematicallysubstitutesthereferencepointersfortheredundantblocks;alsoknownasstoragecapacityoptimization.Dirtydataisdefinedinvariouscategoriesperformancedegradation-asadditionaluselessdatademandmoreprocessing,moretimeisrequiredtoansweruserqueries;qualitylossthepresenceofreplicasandotherinconsistenciesleadstodistortionsinreports andmisleadingconclusions basedontheexisting data;increasingoperationalcostsbecauseoftheadditionalvolumeofuselessdata,extracostsarerequiredonmorestoragemediaandextracomputationalprocessingpowertokeeptheresponsetimelevelsacceptable.Toavoidtheseproblems,itisnecessarytostudythecausesofdirtydatainrepositories.Amajorrootisthepresenceofduplicates,ornearduplicates

intheserepositories,thoseconstructedbytheaggregationorintegrationofdistinctdatasources.Theissueofdetectingandremovingduplicateentriesinarepositoryisgenerallyknownasrecorddeduplication.Inourprojecttoremove thedirtydateinthebloodbankmanagementsystem.Asapartofgeneticprogrammingapproachthegainingconceptsandtheentropycalculationsareusedtodeduplicatetherecords.

II. LITERATURESURVEY

AliteraturehasbeendedicatedtotheRecordDeduplica-tionandtremendousprogresshasbeenmaderangingfromefficient andscalablealgorithmRecordDeduplicationinCoraandRestaurantdataset.Recorddeduplicationisagrowingresearchtopicindatabaseandrelatedfieldssuchasdigitallibraries.Thisproblemarisesmainlywhendataarecollectedfromdisparatesourcesusingdifferentinformationdescriptionstyles and metadata standards. Common place for replicasarefoundindatarepositoriescreatedfromOCRdocuments.Thesituationscanleadtoinconsistenciesthatmayaffectmanysystemssuchasthosethatdependonsearchingandminingtasks.Tosolvetheseinconsistenciesitisnecessarytodesignadeduplicationfunctionthatcombinestheinformationavailableinthedatarepositoriesinordertoidentifywhetherapairofrecordentriesreferstothereal-

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

worldentity.Inrealmofbibliographiccitations,forinstance,thisproblemwasdiscussedbyLawrenceetal.[19],[20].Lawrenceproposedanumberofalgorithmsformatchingcitationsfromdifferentsourcesbasedonwordmatching,phrasematching,andfieldextraction.Asmorestrategiesforextractingdisparatepiecesofevidencebecomeavailable,Elmagarmidetal.[21]followingtwocategories:

A. Ad-HocorDomainKnowledgeApproaches

Thiscategoryincludesapproachesthatusuallydependonspecific domainknowledgeorspecificstringdistancemetrics.Techniquesmakeuseofdeclarativelanguages[21]canbealsoclassifiedinthiscategory.

Training-basedApproaches

Thiscategoryincludesallapproachesthatdependonsomesortoftrainingsupervisedorsemi-supervisedinordertoidentifythereplicas.Machinelearningapproachesfallintothiscategory.Next,brieflycommentonsomeworksbasedonthese two approaches(domain knowledge andtraining-based),thosethatexploitthedomainknowledgeandthosethatarebasedonprobabilisticandmachinelearningtechniques.

Pros-Closestmatchingapproach.

Cons-

1.Forrecordmatchinghighweighttokensarerequired.2.Identification of recordreplication isdone onindividualbasis.

B. Probabilistic Approaches

W.W.Cohen(2002)wasproposed,onestoaddresstherecordduplicationproblemasaBayesianinferenceproblem(aprobabilisticproblem) and proposed the first approachtoautomaticallyhandlereplicasandrecorddeduplicationproblemasaBayesianinferenceproblem(aprobabilisticproblem) and proposed thefirst approach to automaticallymanagededuplication.

ToElaboratedstatisticalmethod.Work,FellegiandSunter[5]proposedanelaboratedstatisticalapproachtodealwiththeproblemofevidence.Thesemethodsrelyonthedefinitionoftwoboundaryvaluesthatareusedtoclassifyapairofrecordsasbeingduplicaterecordsornot.

Febrile[2]toimplementToolsforthismethod,suchas,workwithtwoboundariesasfollows:

Pros-Thepositiveidentificationboundaryisthesimilarvalueliesabovetheboundary,Therecordsareconsideredasreplicas

Cons-

Thenegativeidentificationboundaryisthesimilarvalueliesbelowthisboundary,Therecordsareconsideredasnotduplicaterecords

C. MachineLearningApproaches

M.Bilenko(2003)useamachinelearningtechniquetoim-prove both the similarityfunctions that areapplied to comparerecordfieldsandthewaythepiecesofevidencearecombined.Intheir system, called Marlin, [21][22]. The training datasetisassumedtohavesimilarcharacteristicstotestdataset,which makessolutionusingmachinelearningtechniquesandtogeneralizetheirsolutionsforduplicaterecord.

Pros-UsingSVMClassifierforbettermentofreplicaidentification.

Cons-

ActiveAtlas[28]systemrequiredforrecordmappinginordertoestablishrelationship.

III. IMPLEMENTATIONDETAIL

1) Initializeandsetthepopulation(randomoruserprovidedindividuals).
2) Evaluateallindividualsinthecurrentpopulation,applyanumericratingorfitnessvaluetoeachone.
3) Iftheterminationcriterionisfulfilled,thenexecutethelaststepandcontinue.
4) Reproducethebestnindividualsinthenextgenerationpopulation.
5) Selectn individuals that willprocessthe next generationwiththebestparents.
6) Applythegeneticoperationstoallindividualsselectedandcomposethenextpopulation.Replacetheexistinggenerationofthegeneratedpopulationandgobacktostep2
7) Presentbestindividualsinthepopulationastheoutputoftheevolutionaryprocess.

A.MathematicalModel

Systemforlocatingrecorddeduplicationusinggeneticprogramingapproachillustratethemathematicalmodelthatcontainsduplicaterecords.ByusingGemeticPrograming-cosinesimilaritycanbederivedfromthisdatasetanddedupli-catefunctionsareformedanditisbeingconstructedintheformoftree.

Dataset(ds):-AnyDatasetconataingdifferentrecordfrom differentdomainExtraction(fe):-ExtractFeaturefrominputdataset SimilarityFunction(sf):-FindtheCosineSimilaritybetween Records.

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

TreeConstruction(tc):-deduplicationFunction
$S=\{P,FE,SF,TC,GO,O\}$
$LetD$: $\{d1,d2,.....d_n\}$ whereDconsistsofnoofrecords$LetFE$: $\{f1,f2,f3.....f_n\}$ whereFEisfunctionwhichextractfeatures
$LetSF$: $\{v1,v2,v3,.....v_n\}$ wherevconsistsofsimilarityfunctionrecords
$LetTC$: $\{n1,n2,.....n_n\}$ whereteisatreeconstructorafunctionwhichconstructsatree
$LetGO$: $\{r1,r2,.....r_n\}$ whereGOfunctionisusedtogeneratebestfitnessresult
$LetO$: $\{o1,o2,....o_n\}$ whereOconsistofoutput
FunctionF1returnsthefeaturesextractedfromdatasetrecord
$F1(d)\longrightarrow FE$ fordownloadeddataset
$F1(f)\longrightarrow\{f1, f2,....f_n\}$ εtoFE
FunctionF2returnsthesimilarityvaluesofeachrecord
$F2(FE)\longrightarrow SF$
$e.g.F2(FE)\longrightarrow\{v1,v2,....v_n\}$ εtoSF.FunctionF3returnstheconstructedtreeornode.
$F3(SF)\longrightarrow TC$
$e.g.F3(SF)\longrightarrow\{n1,n2,.....n\}$ εtoTC. FunctionF4returns thegeneticoperationresult.$F4(SF)\longrightarrow GO$
$e.g.F4(SF)\longrightarrow\{r1,r2,...r_n\}$ εtoGO.FunctionF5returnstheoutput
$F5(GO)\longrightarrow O$
$e.g.F5(GO)\longrightarrow\{o1,o2,....o_n\}$ εto O.
FunctionalDependencyoftheabovefunctions

	F1	F2	F3	F4	F5
F1	1	0	0	0	0
F2	1	1	0	0	0
F3	0	1	1	0	0
F4	0	0	1	1	0
F5	0	0	0	1	1

I. SYSTEMDESIGN

A. DataPreprocessing

Cosinesimilarityisameasureofsimilaritybetweentwovectors ofaninnerproductspacethatmeasuresthecosineoftheanglebetweenthem.Thecosinesimilarityvaluof0is1,andlessthan1foranyangle;theminimumvalueofthecosineis-1.Thecosineanglebetweentwovectorsisdeterminestheirtwovectorsarepointinginroughlythesamedirection.cosinesimilaritybetweentwostringsiscalculatedusingLevensteindistanceandSoft-TFIDFsimilaritymethod.

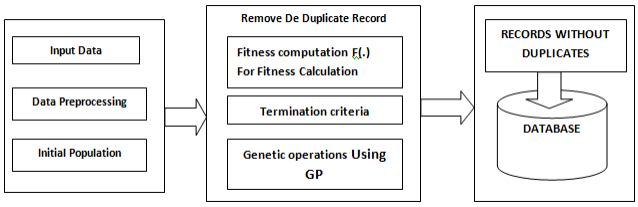

Fig.1.SystemArchitecture

ThecosinesimilarityfunctionusingGeneticprogrammingapproachautomaticallyselectpopulations.Cosinesimilarityfunctionalsousedtointrainingphasetocapturethecharacteristicsofdataset.Basedonthecharacteristicofdatasetonlypopulationsareselected.

B. FeatureVectorExtraction

populationsisthefeaturevectorsselectedbasedoncosinesimilarityfunctions.Iffitnessfunctiondoesnotreachfitnessvaluethenpopulationsarechanged.e.g:(att1,att2,att3)Selectedfitnessfunctionbymachinelearningapproachhastoreachthefitnessvalue.

IffitnessfunctiondoesnotreachfitnessvaluethenhavetochangethefitnessfunctionusinggeneticoperationsGeneticoperationsismutation,crossover,reproduction.Selectedfitnessfunctionhadrepresentedintreeformat,forapplyinggeneticoperationseasily.

C. GeneticOperations

Ifthefunctiondoesnotreachafitnessvaluehavetoapplygeneticoperationsagaintochangethefitnessfunction.

D. Reproduction

ReproductionistheoperationofGeneticprogrammingthatcopiesindividualswithoutmodifyingthem.Theoperatorisusedtoimplementanelitiststrategythatisadoptedtokeepthegeneticcodeofthefittestindividualsacrossthechangesinthegenerations.Ifagoodindividualisfoundinpreviousgenerations,itwillnotbelostwhileexecutingtheevolutionaryprocess.

E. Crossover

Thecrossoveroperationisallowsgeneticcontentwhicharetheprocessofexchangingtwoparents,inthisprocessthatcangeneratetwoormorechildren.InaGeneticprogrammingevolutionaryprocess,twoparenttreesareselectedaccordingtoamatching(orpairing)policyand,then,arandomsubtreeisselectedineachparent.

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

Mutation

Themutationoperationhasbeenimplementedforkeepingamin imumdiversitylevelofindividualsinthepopulationthusitavoidpre matureconvergence.

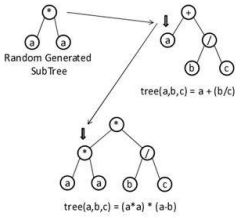

Fig.2.RandomSubtreeMutation

F. RemovingDuplicateRecords

Afterselectingafitnessfunctionfirstlycalculatefitnessvaluef orallrecordsusingafitnessfunction.TheFitnessFunctionselecte dusingexistingmachinelearningapproach.Theselectedfitnessf unctioncanremovetherecords.Beforestoringrecordsintoadatab asecalculatefitnessvaluesforallrecordsiftworecordsmatchwith samefitnessvaluethenremoveonerecord.

II. GENETICALGORITHM

Inproductionscheduling,populationofsolutionsconsistsofm anyanswersthatmayhavedifferentsometimesconflictingobjecti ves.Itoperatesonapopulationofsolutionsratherthana single solution. Forexample, in one solutionit may beoptimizingaproductionprocesstobecompletedinaminimumt ime.Inanothersolutionitmayoptimizeforaminimumdefects.By crankingupthespeedatwhichitproducerunintoanincreaseindef ectsinourfinalproduct.Asincreaseinthenumberofobjectivessys temaretryingtoachieveandalsoincreasethenumberofconstraint sontheproblemandsimilarlyincreasethecomplexity.Geneticpr ogrammingarebetterforthesetypesofproblemswherethesearch spaceismorelargeandthenumberoffeasiblesolutionsisverysmal l.ModelingtheRecordDeduplicationProblemusingGPwhileusi ngGeneticProgrammingtosolveaproblem,thereare someordinary requirementsthat mustbe fulfilled,order tosuccessfullyexplorethistechniquewhicharebasedonthedatast ructureusedtorepresentthesolutionofrecorddeduplication.itch osenatreebasedGPrepresentationfortheevidencecombinationf unction,sinceitisanaturalrepresentationforthistypeofdefinedfu nctions.

Theserequirementsarethefollowing:

1) Theproblemsolutionmustbemodeledasatreestruc-ture.
2) Theevolutionaryoperationsappliedoverthemodeledtreem ustbeattheendoftheresultandconvertedintoavalidtreefinall y.
3) Themodeledtreemustbeautomaticallyevaluatedinorderto maketheuseofthistechniqueviable.

EvidenceEisapair¡attribute,similarityfunction¿thatrep-resentstheuseofaspecificsimilarityfunctionoverthevaluesofaspe cificattributefoundinthedatarepositoriesbeinganalyzed.e.g,itwa nttodeduplicateadatabasetablewithfourattributes(e.g.,name,sur name,address,andpostalcode)usingaspecificsimilarityfunction(e.g.,theJarofunction[Koudasetal.2006]),wouldhavethefollowin gpiecesofevidence:

Algorithm1Algorithm1:GeneticProgramming

Letdatabase(DB)isthesetofdeduplicaterecords;
Letsim(Similarity)isthesimilarityfunctionofDuplicaterecords;
/*DataPreprocessing*/
Generateasetofpairs$P=p_1:::p_n$
fromDB;Computesim(p)foreachpair$p\varepsilon P$;
$R\leftarrow rank P$accordingtosim(p)values;LetTbeth
etopkpairsinR;
LetBbethebottomkpairsinR;
Userlabelseachpair$p\varepsilon T\cup B$with$L_p\varepsilon\{T;F\}$;ComputeaweightW
Pforeach $p\varepsilon P$;
/*ProcessEvaluation*/
$Gen_0\leftarrow$Generatemfunctions;
$Evaluate(Gen_0, T\cup B)$;
Computeaweight$W_f$foreachfunction$f\varepsilon Gen_0$;
$Com_0\leftarrow\varphi$;i=0toapredefinednumberofgenerationsdo
$Com_i top$CfunctionsinGen_i;
/*ActiveLearning*/eachpairp Pdo
Com_ilabelspwith$L_p\{T;F;D\}$;
$if L_p=D$thenUserlabelspwithL' $p\varepsilon\{T;F\}$
/*Reinforcement*/
$UpdateWP foreach p\varepsilon P$;
$Gen_{i+1}\leftarrow$CrossoverMutation$(Gen_i Com_i)$;
$Gen_{i+1}\leftarrow Gen_{i+1}\cup Com_i; Evaluate(Gen_{i+1}, P)$;
Compute$W_f$foreachfunction$f\varepsilon Gen_{i+1}$;
$Com_i\leftarrow$topCfunctionsinGen_i;
foreachpairpεPdodo
Com_ilabelspwith$L_p\varepsilon\{T;F\}$

E1(name,Jaro),E2(surname,Jaro),E3(address,Jaro),andE4(postal code,Jaro)Forthisexample,averysimplefunctionisalinearcombina tionsuchasFs(E1;E2;E3;E4)=E1+E2 +E3+E4.

I. EXPERIMENTS

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

A.Dataset

Apartfromthetasksrelatedtothelogicalandthephysicalsideofthedatabasesysten,DatabaseAdnibistratormayalsotakepartindatabaseSystemoperations.mainroleistogivedevelopersrecommendationsabouttheDBMSspecificities,thushelpingthemavoidanydatultingimabaseperformanceissues.OtherimportantworkoftheDBAsarerelated todatamodelingatoptimizingthesystem,aswellastothecreationanddanalysis of newdatabases. computer science,isevolutionarycomputationsubfieldofartificialintelligence(moreparticularlycomputationalintelligence)thatinvolvescombinatorialoptimizationproblems.

TousedRestaurantandCoradatasetstoanalyzetheproposedalgorithmandtheperformanceoftheproposedalgorithmiscomparedagainstthegeneticprogrammingtechniquewithtthehelpofevaluationmetricsInthisexperimentselectdatasetsfromthecoradatasetdatarepositoryandthedatasets usedisRestaurantdataset.Thedatasets,whichareusedinproposed approach,isdetailedbelow.Dataset1(CORA):TheCoradatasetconsistsofduplicateandnon-duplicatesdatarecordswhichiscitedto122Conferencepaper,itisdivide intovariousattributes(name,year, title, and other information)Dataset2(Restaurant):Thisdatasetcontains500 records(400originalsand100duplicates),duplicatesrecordbased ononeoriginalrecord(usingaPoissondistributionofduplicate records) andwitha maximumoftwomodifications inasingleattributeandinthefullrecord.

I. CONCLUSIONANDFUTUREWORK

Endingwiththeduplicateinformationisstoredfromdifferentsourcesthatrequirestoragespaceforstoringreplicasorded uplicaterecords.severaleventhasbeenexperimentedon modem enterprise and data loss wascommon in all .Variousmethodshavebeentriedtotacklewiththatproblemsident ificationandreplicashandlingisimportantforguarantingthequality ofinformationwhichisbeingavailablebydataintensivesystem suchasdigitallibrariesande-commercebrokers.Thisprojecthelpinpresentingsystemforlocating recorddeduplicationviageneticprogramming.Ourapproach iscombinationofseveralpiecesofevidencethathasbeenextracted fromthedatacontentfortheproductionofdeduplicationfunctionthatenablesforidentificationwhethertwoormore entriesareintherepositoryorreplicasornot.Infuturethisapproach willhelpinfindingthecomplexmatchesinthreedifferentdatarepositoryscenarioswheretherepositoriespartiallysharesomedatarsomecommondata.Theentireproposedworkconceptisconcluded withtheliteraturesurvey ,designandmodelling,projectobjectiveusingsystem.Forlocating recorddeduplication.Ourevolutionaryapproachhasimplemented rangeofuseforourfuturework.Experimentswerecarriedoutforevaluationourworkwithdatafromspecificapplicationdamainswithdifferentdatasets.Otherinvestigationtechnigueswerecarriedout.Thispapersuggestsefficientimplementationofk-NNforbettermento fresultsandadvancementintechnology.

ACKNOWLEDGMENT

IamthankfultoProf.Dr.V.H.PatilHODandvicePrincipal,MCERC,Nashikforgivingherprecioustimeandguidelineduringthispaperalsoforherexpertguidanceandcontinuousencouragementthroughoutthispaper.Iwouldliketoexpressdeepestappreciationtowards Prof.Dr.G.K.Kharate,PrincipalMCERC,Nashik,andProf.RanjitGawande(ME Coordinator) whoseinvaluable guidance supported me incompletingthispaper.

REFERENCES

[1]oisesG.deCarvalho,AlbertoH.F.Laender,MarcosAndreGoncalves,andAltigranS.daSilva,AGeneticProgrammingApproachtoRecordDeduplicationIEEE TransactiononknowledgeandDataEngineeringvol.24,No,.3,March 2012

[2].Agrawal,S.Chaudhuri,G.Das,A.Gionis,Automatedrankingofdatabasequery results,in:ProceedingsoftheFirstBiennialConferenceonInnovativeDataSystemResearch,2003

[3].Koudas,S.Sarawagi,andD.Srivastava,"RecordLinkage:SimilarityMeasures andAlgorithms,"Proc.ACMSIGMODInt'lConf.ManagementofData,pp.802-803,2006.

[4].BhattacharyaandL.Getoor,IterativeRecordLinkageforCleaningandIntegration,Proc.NinthACMSIGMODWorkshopResearchIssuesinDataMiningand KnowledgeDiscovery,pp.11-18,2004.

[5] .P.FellegiandA.B. Sunter,ATheory for RecordLinkage,J. Am. StatisticalAssoc.,vol.66,no.1,pp.1183-1210,1969.

[6].S.Verykios,G.V.Moustakides,andM.G.Elfeky,"ABayesianDecisionModelfor CostOptimalRecordMatching,"TheVeryLargeDatabasesJ.,vol.12,no.1,pp. 28-40,2003.

[7].BellandF.Dravis,"IsYouDataDirty?andDoesthatMatter?,AccentureWhiterPaper,http:/www.accenture.com,2006

[8] R.Koza,GenticProgramming:OntheProgrammingofComputersbyMeansof NaturalSelection.MITPress,1992.

[9] .M.deAlmeida,M.A.Gonalves,M.Cristo,andP.Calado,ACombinedComponent Approachfor FindingCollection-AdaptedRankingFunctionsBasedonGeneticProgramming, Proc.30thAnn. Int'l ACMSIGIRConf.ResearchandDevelopmentinInformationRetrieval,pp.399-406,2007.

[10] Zhang,Y.Chen,W.Fan, E.A. Fox,M.Gonalves, M.Cristo,andP.Calado,IntelligentgpFusionfromMultipleSourcesforTextClassificationProc.14thACMInt'lConf.InformationandKnowledgeManagementpp.477-484,2005

[11]G.deCarvalho,M.A.Gonalves,A.H.F.Laender,andA.S.daSilvaLearningtoDeduplicate,Proc.SixthACM/IEEECSJointConf.DigitalLibraries,pp.41-50,2006.

[12].BilenkoandR.J.Mooney,AdaptiveDuplicateDetectionUsingLearn-

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

ableStringSimilarityMeasures,Proc.NinthACMSIGKDDInt'lConf.KnowledgeDiscoveryandDataMining,pp.39-48,2003.

[13] Lawrence,L.Giles,andK.Bollacker,DigitalLibrariesandAutonomousCitationIndexing,Computer,vol.32,no.6,pp.67-71,June1999.

[14].K.Elmagarmid,P.G.Ipeirotis,andV.S.Verykios,DuplicateRecordDetection:A Survey,IEEETrans.KnowledgeandDataEng.,vol.19,no.1,pp.1-16,Jan.2007

[15] W.Cohen,DataIntegrationUsingSimilarityJoinsandaWord-BasedInformationRepresentationLanguage,ACMTrans.InformationSystems,vol.18,no.3,pp.288-321,2000

[16]C.P.CarvalhoandA.S.daSilva,"FindingSimilarIdentitiesamongObjectsfromMultipleWebSources,"Proc.FifthACMInt'lWorkshopWebInformationandDataManagement,pp.90-93,2003.

[17] B.Newcombe,J.M.Kennedy,S.Axford,andA.James,AutomaticLinkageofVitalRecords,Science,vol.130,no.3381,pp.954-959,Oct.1959.

[18] FreelyExtensibleBiomedicalRecordLinkage,http://source-forge.net/projectsfebrl,2011

[19].W.CohenandJ.Richman,LearningtoMatchandClusterLargeHigh-DimensionalDataSetsforDataIntegration,Proc.NinthACMSIGKDDInt'lConf.KnowledgeDiscoveryandDataMining,pp.475-480,2002

.

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015

'FUTURIZM-2014, (14TH& 15TH March 2014), 'P.D.E.As College of Engineering, Manjari[Bk]', 'IIP Cell,'Manjari Road, Hadapsar, Pune-412307. India.',Chief Editor:Dr.K.R.Harne', Editors: Prof R V Patil, Prof Niraja Jain

International Journal of Innovative Technology & Adaptive Management (IJITAM)
ISSN: 2347-3622, Volume-2, Issue-7, April-2015