EXPERIENCE MANAGEMENT FOR IT MANAGEMENT SUPPORT

by

Can Bozdogan

Submitted in partial fulfillment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
July 2012

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “EXPERIENCE MANAGEMENT FOR IT MANAGEMENT SUPPORT” by Can Bozdogan in partial fulfillment of the requirements for the degree of Master of Computer Science.

Dated: July 30, 2012

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: July 30, 2012

AUTHOR: Can Bozdogan

TITLE: EXPERIENCE MANAGEMENT FOR IT MANAGEMENT SUPPORT

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: M.C.Sc. CONVOCATION: October YEAR: 2012

_____

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

This thesis focuses on the identification of experience required for solving IT (Information Technology) problems in small to medium sized enterprises. It is aimed to utilize information retrieval and data mining techniques to automatically extract information from publicly available experience data on the internet to automatically generate a knowledgebase for dynamic IT management support. In this thesis, similarity distance measures as Jaccard Index, Cosine Similarity Measure and clustering algorithms as K-Means, EM, DBScan, CES, CES+ are employed on three different datasets to evaluate their performances. CES+ algorithm gives the highest performance results in these evaluations. Moreover, Multi Objective Genetic Algorithm (MOGA) is used and is evaluated on three different data sets to aid the usage of CES+ in real life senarios by automating the selection of necessary parameters. Results show that MOGA support is not only automating the CES+, it also provides higher performance results.

List of Abbreviations and Symbols Used

NSERC	Natural Science and Engineering Research Council of Canada
FGS	Faculty of Graduate Studies
FCS	Faculty of Computer Science
NIMS	Network Information Management and Security Lab
IT	Information Technology
IR	Information Retrieval
FAQ	Frequently Asked Questions
MOGA	Multi Objective Genetic Algorithms
BDIM	Business Driven Information Management
UI	User Interface
PIHK	Princeton IT helpdesk knowledgebase web site
tf	Term frequency
idf	Inverse document frequency
EM	Expectation Maximization Algorithm
p()	Conditional probability
θ	Given condition
$\propto$	Content value
CosSim()	Cosine similarity measure
$\overrightarrow{tfıdf}$	Feature vector of each cluster for CES algorithm
w_i	i th word of a dictionary of a cluster for CES algorithm
$\cap$	Intersection
$\cup$	Union

$\in$	Element of
$\sum$	Array summation
GA	Genetic Algorithm
NSGA	Nondominated Sorting Genetic Algorithm
NPGA	Niched Pareto Genetic Algorithm
C	Content value representation of CES+MOGA
D	Dictionary threshold
CC	Creating Core clusters threshold
EC	Expanding the Clusters threshold
S	Sophisticating the Clusters threshold
$Words_{CQ}$	Occurrence of a word in the question part of a data instance
$Words_{CA}$	Occurrence of a word in the answer part of a data instance

Acknowledgements

This research is supported by a Natural Science and Engineering Research Council of Canada (NSERC) Strategic Project Grant. My thanks to FGS/FCS for the partial funding for my Master of Computer Science Degree and my thanks to my supervisor Dr. Nur Zincir-Heywood for all her time, guidance and support.

This work is conducted as part of the Dalhousie NIMS Laboratory, http://www.cs.dal.ca/projectx/.

Chapter 1 - INTRODUCTION

Computer networks and systems are vital components of many organizations today. They are the key component for stocking information, reaching that information and keeping the information flow within the company as well as with the outside of the company. All sizes of companies in all areas of private sector or government install these systems to provide their services or goods in more efficient ways to their customers, and to provide a more efficient environment for their employees. Even one of the first concerns of a new startup company is implementing IT structure plan that will be used for the business model.

Creating an IT infrastructure plan and building it up is just the beginning for a company and an organization. After building it up, the biggest concern is keeping it running with minimum business interruptions. Providing 24/7 services is majorly related to having a 24/7 available IT infrastructure within the organization to keep internal and external processes going and to keep the organization's systems and network up and running. To provide efficiently running IT systems, organizations follow a discipline to manage all their technology resources in accordance with their needs and priorities. This is called as IT management. Failures in IT systems are unavoidable and because of that keeping the IT system on service without interruption is provided by minimizing the correction time of failures in the system. Fast correction of any IT related network or server problem/failure is required to minimize business interruptions to provide efficient service.

Correcting network and server failures is a time consuming and knowledge intensive process in practice. In depth knowledge of the network/system setup and the relevant monitoring tools is required to diagnose the problem and to perform necessary corrective actions quickly. One way to gain knowledge, to improve skills for diagnosing and correcting problem/failure corrections for IT management team members is using available information about the system and problems/failures. In practice, IT team members accumulate experience over time by using trouble ticket systems, diagnostic reports, application log files etc. However, providing a team member's own knowledge to other members in the team or possible new members of the team is time consuming.

Knowledge on historical problem/failure cases and their corrections are called as experience. When a problem occurs, most of the troubleshooting time is spent on two steps of the trouble shooting process as: (i) Identification of the root cause and (ii) Planning the resolution. A solution to reduce the troubleshooting time is a crucial task because of the concepts that are mentioned before. Taking advantage of gained experience and retrieving previous cases similar to a current problem case is a good way of providing support for the system/network administrator to apply past knowledge to develop corrective actions for the system problems/failures. Such a system may lead to the second generation troubleshooting support systems to manage the availability of edge network and system components. Indeed, first concern of such a system is generating a knowledgebase as well as the structured representation of cases and efficient semantic similarity metrics for cases from the available experience. Additionally, despite being an effective solution for fast correction of system problems, such a system faces another issue as having lack of experience data to generate a knowledgebase. In big sized companies because of accumulated experience over many years by many IT team members, source of experience data to generate a knowledgebase is not a big deal. However in small to medium sized companies or an organization that just had a start up, lack of experience data is a problem for an IT management support system.

One way of solving the above issues is following an information retrieval (IR) approach to take advantage of publicly available historical experience for IT management support. Thus, the goal is to utilize information retrieval and data mining techniques to automatically extract information from publicly available resources such as FAQs (Frequently Asked Questions), forums, IT help web pages in order to automatically generate a knowledgebase for dynamic IT management support. Such a system can save resources such as time, cost etc. especially for small to medium sized organizations and companies where having experience data and having experienced IT personal is an issue. Additionally, such a system can be useable for all size of organizations when they decide to use a new IT structure by changing their current structure. Change in the current structure can have different aims such as to have more secure or effective or cheaper information flow control within and with outside of the organization. In this case, it is possible for the IT management team members not being familiar with that new

technology. For example, consider a company that uses enterprise resource planning software internally on its own server. And consider that after years, management team of the company decides to use cloud and distributed computing systems instead of using the internal server centric system for their enterprise resource planning. In this case, if IT management team of the company does not have enough knowledge and experience about the cloud and distributed computing systems, fixing problems/failures of the system will be a hard task, which takes so much time. However, by crawling useful data about cloud and distributed computing system from the publicly available internet resources and generating a database, they can get support and minimize the time for fixing problems/failures by using the proposed system.

Thus, the main focus of this thesis is providing the right mix of automation and manual activity to minimize the overall expense of the IT and network management tasks.

In this thesis, two similarity distance measures as Jaccard Index, Cosine Similarity Measure and five clustering algorithms as K-Means, EM, DBScan, CES, CES+ are employed on different datasets to evaluate their performances. In the experiments three different data sets are used. These are obtained from publicly available FAQ knowledgebases of well known helpdesk web sites. Additionally, the performance of MOGA is evaluated to give automation support for the CES+ algorithm. In a real time real life applications, it is very important to use an algorithm that can automatically generate its parameters on optimum level to deal with different data sets. Same data sets are used with the cross validation method to provide a comparative environment for the evaluation of manually configured CES+ and automatically configured CES+ by MOGA.

Thus, the first part of this research is to investigate whether IR approach with a clustering algorithm is able to perform well on generating an experience knowledgebase from publicly available experience data on the Internet to provide support for IT management tasks. The second part of this research is to investigate whether it is possible to provide automated parameter optimization for the clustering algorithms evaluated by using MOGA to be able to use the clustering algorithms with minimum human intervention on real life data sets. In this research, CES+ algorithm gives the best results in terms of extracting the most similar past experiences where an experience instance is defined as a problem and its solution. Results show that providing MOGA support is not

only automating the CES+, it also provides higher performance results in terms of precision, recall, *Fwithin* and *Fbetween* performance metrics. Finally, a basic software is implemented to make the proposed system available for testing in a more user friendly environment.

In the following chapter 2, Literature Survey, summarizes the existing works in the literature where different IT and network management support systems are developed with different approaches. Chapter 3, Methodology, presents the learning algorithms that are employed in this research; basic description of the learning algorithms, datasets employed for training and testing, the preprocessing of the datasets, the proposed model's architecture, and the software tool that is developed to provide an easy usage of the proposed system. Chapter 4, Experiments and Results, describes the evaluation experiments performed and the metrics that are used to evaluate performances of the algorithms employed. Chapter 5, Conclusions and Future Work, draws conclusions and gives directions for the future work.

Chapter 2 – LITERATURE SURVEY

In IT management, the key problems that consume IT management teams' time are diagnosing system/network states and correcting problems. Thus, research in this area is really ambitious to find ways of reducing time and cost caused by these problems for IT management teams in organizations and companies. Researchers aim to fully automate the probing of network such as in [1] or perform ontology mapping for interoperability in network management such as in [2]. Additionally, representation of experience has been typically addressed as a case-based reasoning (CBR) problem [3] in the literature. Cases have been represented as free text descriptions, conversations or structural descriptions. Structural descriptions are the most relevant to the problem presented in this thesis and they have been classified into attribute-value, graph, object-oriented and predicate logic representations. In [4], Iwai et al. use structural description for experience data and CES algorithm to generate heterogeneous representation of experience for the support of IT management teams in organizations. Although it is an efficient algorithm, CES has high computational cost and it has no mechanism for optimum parameter selection to deal with different data sets, which are not seen previously. Moreover, there are different successful approaches in the literature such as incident management [5,6], fault management[7] and using historical data [8] in literature to deal with IT management problems.

Bartolini et al. reviewed business-IT alignment, business-driven IT management (BDIM), and anatomy of a solution for such problems [5]. They applied business driven IT management approach to develop a solution for IT incident management. The division in the IT support team is used to create support levels. From low level to high level, profession of the members of the group is getting more and more specialized. Additionally, higher support groups are determined by their specialized knowledge of the tasks such as servers, network, firewalls etc. Helpdesk is the entrance of an incident for the IT support team in their approach. When a customer has a request, Help Desk creates an incident and assigns a priority level for it to make sure that it is assigned to the right IT support team. In this model basically a set of hierarchical support groups are created, and each group has its own operators. Another important concept in this model is defining

priority level for each incident. The priority level of an incident is defined by calculation through its effect and urgency. This process aims to harmonize needs of the organization with IT by calculating priorities of the incidents and the priority calculation is depending on the business objectives. There are two important concepts for the performance of this model. First one is related with routing of the incidents and called as "effectiveness" [5]. Effectiveness is the metric for defining if an incident is routed to the correct support group. Correct support group means that the group that is equipped to take care on that incident. Second important concept is related with the support groups and it is called as "efficiency" [5]. Efficiency is the metric for defining if a support group is at the right level. By this work, authors review the BDIM literature and employ the above model for incident management to ensure alignment of business and IT.

In [6], Bartolini et al. discuss the decision making in BDIM to provide support for improvement of current tools. The most important decision making problems in BDIM are described as forecast and uncertainty, human decision making and modeling the IT-business linkage. Forecast is an important concept for making decisions on business objectives to obtain better future performance. Another important decision-making problem in BDIM is stated as human decision making. Restructuring IT concepts in organizations to achieve success in business objectives creates the one of the biggest cost for management. Decisions should be made with caution before turning them into practical methods. Because a simple decision, which is considered as correct in an IT model, can result in huge expenses and time wasted if it is not correct for the practical real time business concerns. BDIM tools can support decision making for humans by using different techniques. For example, focusing on the most important variables of the necessary information only, while ignoring the other information with reasonable visualization is one of the opportunities that BDIM tools can support. Bartolini et al. present that machine learning methods represent a promising approach for analyzing human decision making criteria [6]. The last problem which is one of the most challenging problems in BDIM discussed by authors is Modeling the IT-Business Linkage. This means creating a relationship between the performance of an IT structure and the business value that is generated for your objectives by the IT structure. They define two processes as (i) forward, which is used to measure how well defined the

relation between IT function and business objectives is defined, and (ii) backward, which is used to obtain the required changes in the system alignment. Authors evaluated SYMIAN [9] and HANNIBAL [10] decision support tools and they received accurate modeling results. They conclude that the analysis of the forward and backward processes in theoretical model development can support BDIM decision-making.

In [11], Bartolini et al. presents a decision support tool, which is named as Symian-Web. It is a tool that uses what-if scenario analysis to improve IT support. The tool uses cloud system for computing capabilities. Symian is used to create and simulate model of IT organizations for real life and these models are used to see the effect of IT strategies and their processes. The authors use the Symian to capture their incident management model's metrics such as incident resolution and the average number of support groups visited by the incidents. Symian-Web is a tool that creates workflow for performance optimization for IT managers to investigate the model of IT support organization. It also enables user to use features for optimization according to business driven measures. Additionally, it provides a user interface to let users to reconfigure the IT support organization model. To catch business objectives, Symian-Web enables users to compare different configurations of the organization model and this aims to optimize the performance of IT support organization. Performance analysis is done by using reporting functions of the Symian-Web. This allows users to collect the performance data and analyze it. The data is visualized to make it easier for the users. Furthermore, the users can customize the interface by changing the definition of the dimensions of a node. The architecture of the system has layers as the top layer for common presentation functions, the middle layer for main Symian-Web functions, the third layer for providing cloud-based support, and the lowest layer for simulation execution tasks. The Symian-Web is a tool for IT support performance optimization. The authors show that what-if scenario analysis and visualization of the tool is effective in investigating performance in IT support organizations.

Furthermore, in [4], Iawi et al proposes a helpdesk support system with filtering and reusing e-mails. A Help Desk Support System is a system that helps operators at the helpdesk to answer the questions of the users asked via email by using Frequently Asked Questions (FAQ) knowledgebase. This system has two components: (i) replying to

inquiries, and (ii) FAQ. In the first component, the inquiry emails are filtered to define if they are related inquiries to the FAQ. For FAQ related email inquiries, the matching inquiry and question set is retrieved and an automatic reply for the inquiry is sent. For the building up a FAQ step, they propose a clustering method to display suitable candidate FAQs to the help desk operator. In the rest of this thesis, this method is called as the CES clustering method. The clustering method has three basic steps as making core clusters by high threshold value, expanding clusters by low threshold value and sophistication of the clusters. For the testing process of the algorithm, the authors use Jaccard coefficient and Cosine similarity for comparison. As a dataset, they use a sport membership administration Web site's FAQs. As comparison metrics, they use the size of the cluster and the precision of the cluster. Iwai et al. show that their algorithm gives better results on all clusters on their dataset [4].

George et al. proposed an information retrieval based approach [12] to a similar problem. The system aims to assist the system administrators to correct failures. Identification of similar system faults from a knowledgebase is used to utilize the previous failure cases. The authors use term-based vocabulary to represent the problems send by the users and previous knowledgebase cases. They propose a technique that uses a taxonomy based approach for reducing the term space required to represent cases in the knowledgebase. By using a taxonomy based approach, they aim to mitigate the problem of having limited keywords in system events and cases. When a new case is generated, the system uses taxonomy to provide keywords for the initial description. In this case, having an efficient and qualified taxonomy is important to give good support for choosing keywords for the new cases. The taxonomy is a tree structured collection that represents the relationship between different concepts in the dataset. If you traverse up to the root, concepts get more generic, on the other hand if you traverse down to the leaves, concepts get more specific. The authors initially created the taxonomy manually, then they expanded this taxonomy with a combination of automatic selection and manual refinement. They conclude that the small taxonomy that they created manually provides weak representation of their documents but the bigger taxonomy with new sets of candidate terms provides increased coverage. After testing the proposed methodology, the authors conclude that better taxonomy usage can improve the results.

Santos et al. proposes a conceptual system that aims to identify the root causes of the problems [13]. Their conceptual solution reuses the partial or complete cases by support of extension of Common Information Model. Architecture of the system is consisting of four components as input processor, weight calculator, question selector and question verifier. Input processor identifies and lists the factors that may have caused or influenced the failure. After that it creates relations between categories, the application, which had the failure and root cause. Weight calculator prioritizes the root causes, which are related to the failure. This prioritization is done to increase the chance of identifying the correct root cause for that failure. Question selector chooses the questions that are going to be asked to the user by using the weights of the questions that are identified as related to the failure. Firstly the category is defined and then in that category the question is defined depending on their weights. Finally, the question verifier identifies how many times the selected question is asked and identifies the responses that are given by the operators. The answer with the highest weight is assumed as the answer otherwise it is asked to the operator to choose one of the answers. This process heads one to the new questions to define root cause or directly to the root cause. The authors conclude that they received good and effective solutions for the failures.

Xie et al. presented a model to support experience management [14] by organizing development of multiple projects around a single organizational learning and by reuse infrastructure. Experience package lifecycle is presented in four phases as create, store, acquire and reuse. To create an experience package the organization collects the knowledge and then stores it in the experience base. The user can acquire the experience by query and reuse it the experience data. The authors developed a prototype system based on their proposed experience management system which aims to support small organizations to benefit from every expert's experience in the organization. They used three layers as user interface, experience management server and data layer. The user interface is to provide support for the user to use the system efficiently. The server layer is used to map all activities between user interface and data layer for operations such as create, update and search experience packages. The data layer is the main repository, which consists of all experience packages stored. The authors describe a

lightweight approach for experience management task for small organizations to capture, browse and reuse experience from previous projects.

Marcu et al. presented an information model for inter-organizational fault management [8]. Fault management aims to prevent occurring of incidents to avoid their loss. The core objectives of the problem management and the incident management are related to the fault management task. The model proposed by Marcu et al. aims to provide better service quality by more controlled service delivery and support. Three problems as: (i) the outsourcing problem, (ii) the problem of heterogeneity and autonomy, and (iii) the problem of service delivery diversity are defined as specific problems of the interorganization fault management. The authors presented a model to cope with these problems by analyzing phases of fault life cycle to define requirements. In this approach, entities related to the same topic are grouped together and assigned to the same domain. The authors define three domains; (i) Resource that is used to discriminate the resources as intra-organizational and inter-organizational, (i) Service that is used to identify the service with the information about it and its possible problems, (iii) Fault that identifies the fault with the resource alarm and the service failure information. In their work, authors gave an overview of the presented information model's fundamental requirements.

Li et al. proposes a solution to automate the change management process by using past experiences to make more appropriate assessment [7]. Organizations and companies face with change requirements in their IT services for efficiency. But during the changes in organizations, potential service interruptions can result with business loss. To prevent service interruption in the change management process, Li et al. Aim to reduce the time and the cost. They use a business-driven approach for automation of the process by guiding the change by the usage of historical experience. The new requirements, for more effective IT service management brings need for changes to prevent service interruptions. Manual activities cause problems such as reducing the efficiency, increasing the cost etc., because of that for efficient change, automation with previous solutions is the aim to use in the management solution. The model is defined as analyzing the change requirements by extracting information from a natural language description. For better service and efficiency, the authors propose a machine learning method to identify if the changes are

similar to the goals that are defined. The authors mention that they aim to automate the change management process with business-driven perspective by using their proposed machine learning algorithm and they are able to retrieve most similar change solution for service goals.

In summary, in this thesis, the objective is to provide an experience management based IT support system for small to medium size organizations, which may not have big IT support teams or systems. As mentioned before, one of the biggest problems of IT management teams is diagnosing the problems/failures of computer systems and correcting these problems/failures in minimum time to keep business interruptions of organizations and companies minimum. Because of being inexperienced about the current problem and because of not being able to share experiences of IT management team members with each other or with possible new members of the team, solving system problems/failures is a time consuming and a hard task. And to solve this issue, using support tools which aims to benefit from the historical experience, brings another issue as lack of experience data for small to medium sized companies because of not being in the industry for many years and because of low information flow within the company. Thus to this end, this thesis designs and develops an experience management system that employs historical experience data in the form of previously experienced problems and their solutions. Iwai et al. presented a help desk support system by using CES algorithm but it is not practical for a real life system due to computational cost and parameter selection issues. There are different approaches existing in the literature to provide support for help desks such as [12,13] but none of them aims to small to medium sized organization by addressing the lack of employee and historical data issues. Additionally, incident management approaches for providing support to the IT management such as [5,6] can be used in a cooperative way for the proposed system in this thesis.

Chapter 3 – METHODOLOGY

Resolution of network and server problems are among the difficult tasks of IT management. The aim is to provide a system, which gives support for IT management, for small to medium sized organization and companies. The focus in this system is not to perform diagnosis or to correct the problems as in the diagnosis and correction systems [15,16,17]. Here the focus is on retrieving the historic experience to see what is known about the correction of similar faults in the past. Correcting faults automatically is a very difficult task for a computer system but it can be solved by a member of the IT management team who is supported adequately. Such a support can be provided by a system, which inspects structured documentation of successfully solved cases of a similar nature in the past.

Systems such as fault detection, trouble ticket systems, inquiry management and system monitoring are already available as open source or commercial off the shelf. The output of such systems could be the input for the proposed system that has three core components:

(1) The Experience Knowledgebase

(2) The Problem Solution Matching Engine

(3) The Problem Capture Engine

The Experience Knowledgebase consists of previously resolved cases in a structured format. The Problem Solution Matching Engine searches for a match in the experience knowledgebase for the member of IT management team and retrieves the similar cases from the history. The Problem Capture Engine transforms the output of the problem detection systems into an input that can be used by the system. Two major tasks as minimizing experience acquisition efforts and extracting information from experience data are aimed by the system. To minimize experience acquisition efforts the knowledgebase is constructed automatically. The basic structure of the system with its components and experience data flow cycle is shown in Figure 1.

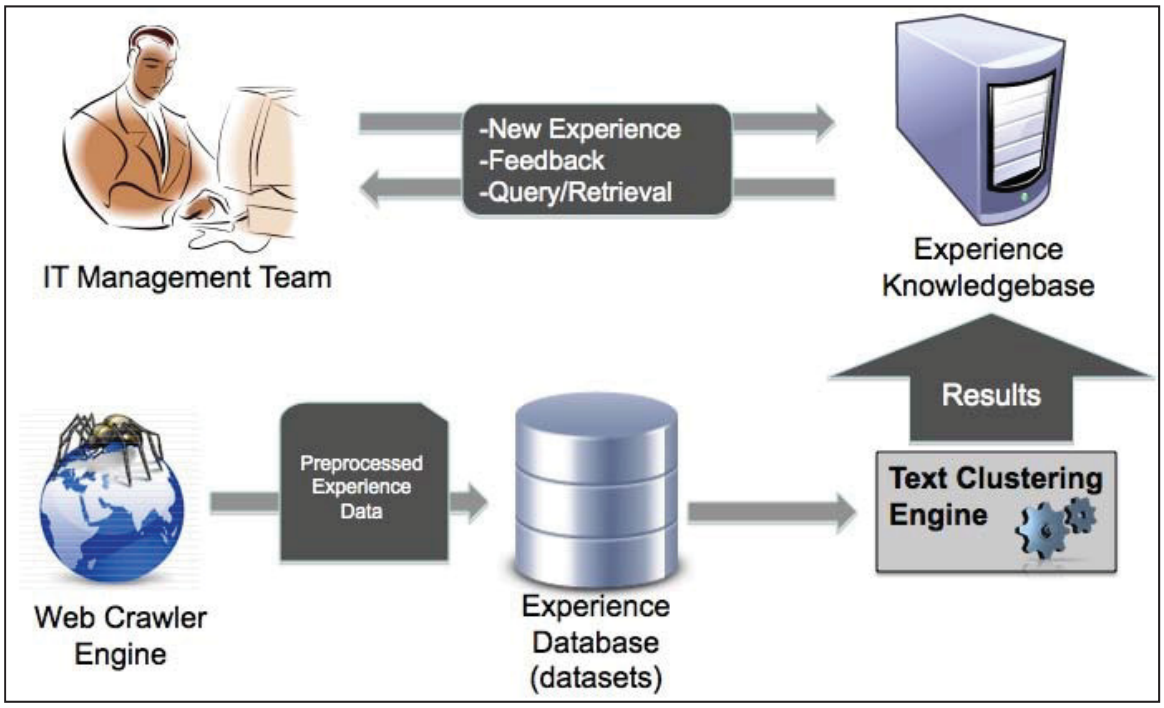


Figure 1: Components and experience data flow of the system

Depending on the components and experience data flow of the system, a software tool is developed with a user interface (UI) to make this system available for real time usage. In the tool, three engines and the experience knowledgebase are generated from the components of the proposed system to make the UI more user friendly and for easier understanding and usage of the system. The software tool consists of three engines:

(1) Experience Data Engine which consists of Web Crawler and the Experience Database

(2) Text Clustering Engine, which consists of the CES+ and the MOGA algorithms

(3) Main window with Problem/Solution Matching Engine, which enables users to send queries to the system and manage other components as well as the Experience Knowledgebase.

The proposed system provides a module for each of the components. To provide better understanding of the system, description of the components are represented in the following. Figure 2 shows the basic structure of the proposed system and the modules that enable the user to manage the components of the system.

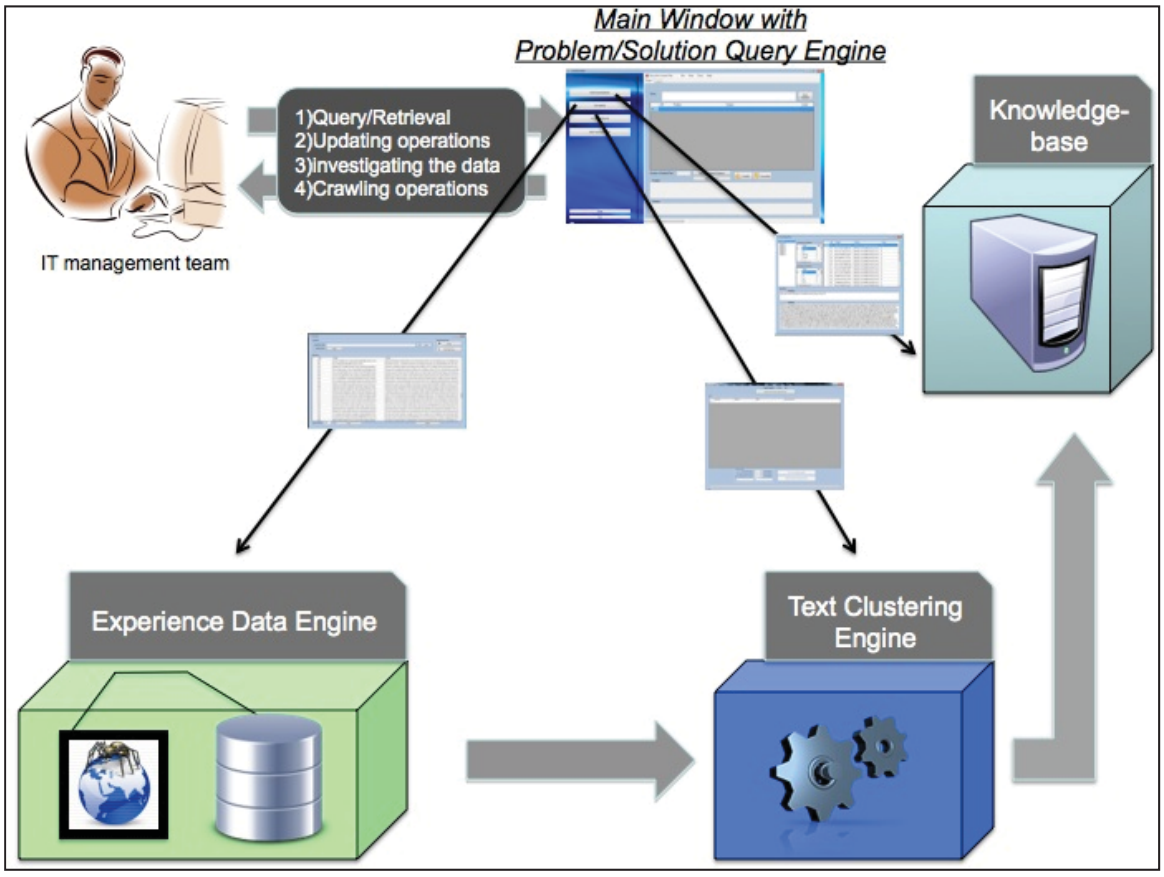


Figure 2: Modules and the architecture of software tool

In the following subsections, the components of the system are described with the background information and with the employed methodologies. These are used to create the system. Experience Data Engine consists of the Web Crawler Engine and the Experience Database.

3.1 Experience Data Engine

Experience Data Engine collects and stores the experience data which is kept for information retrieval and data mining purposes. The engine enables users to view, add, update and delete the experience data instances. Figure 3 shows the Data Base module of the system which is the main window for the Experience Data engine with its relation to the system architecture.

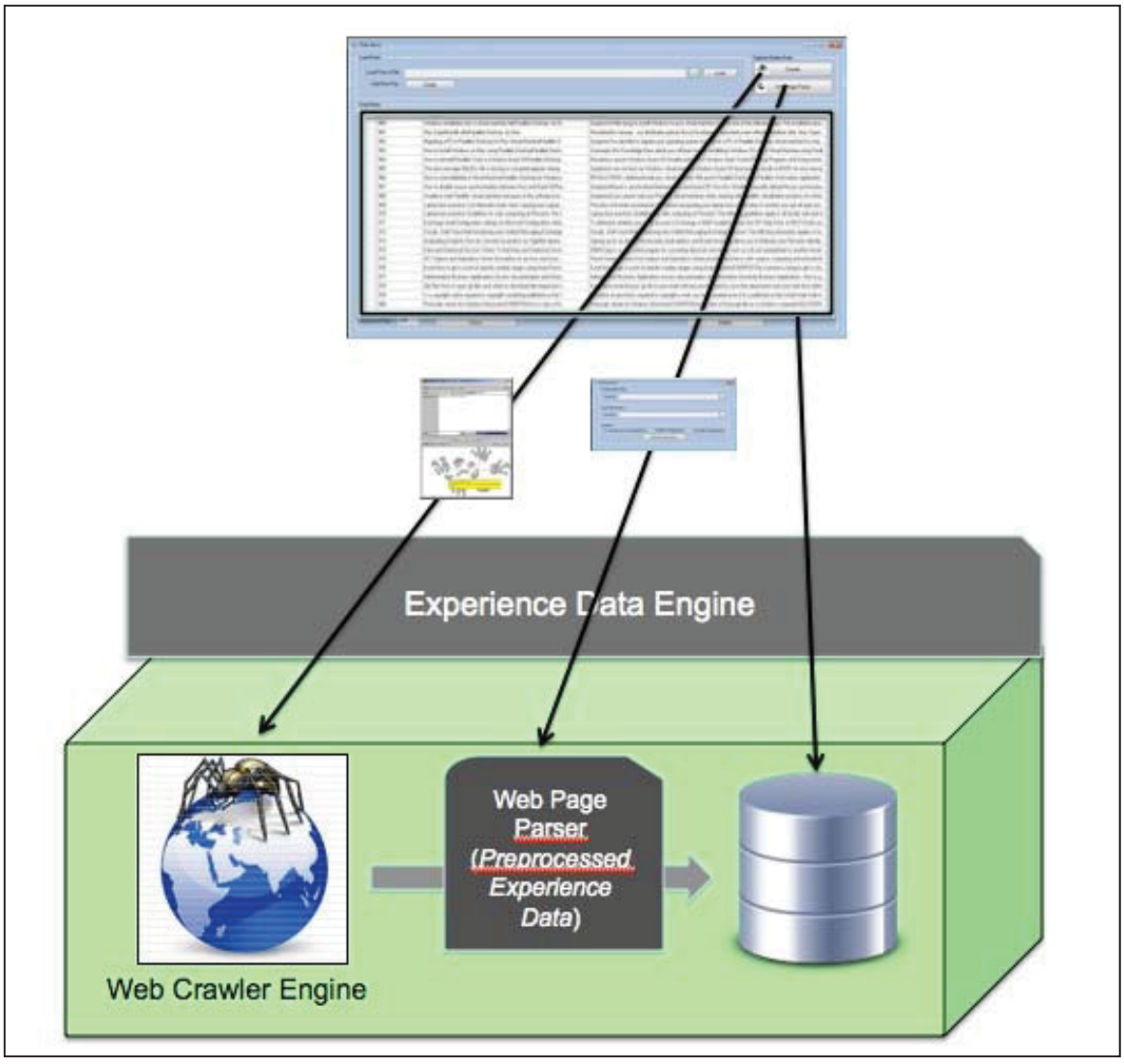


Figure 3: Data Base module of the system and Experience Data Engine

The form window has “LoadData” component, which enables user to load data to the system. It has Data Base table to view the current stored data. A user can investigate the experience data by using the “Focus” button on the screen. This button opens another form that shows contents of the problem and the solution sections of the experience data instance. By using the “Delete” button, the user can delete an experience instance, which is chosen from the database table, from the data set. Additionally, Database module enables the user to gather data from web pages by crawling and then parsing it into problem-solution structure, which can be loaded to the system, by preprocessing the “Capture Online Data” component. Data base module has three sections which are detailed as the following:

3.1.1 Loading Data

To store data into the database, the system provides two ways for the Loading Data component:

(i) Creating a new record for an experience instance with the problem and solution components: In this case, any member of the IT team using this software can create a problem-solution pair based on his/her experience. He/she can use this component and after clicking the "Create" button, a form will appear to simply create a new problem-solution pair as in Figure 4:

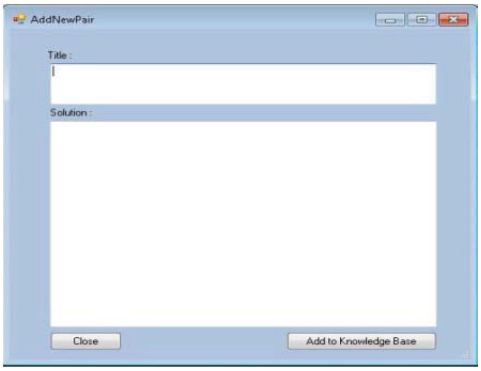


Figure 4: Form window to create your own experience data instance

(ii) Upload an input file which is created after crawling web pages for useful data: To load data from a data file, one can browse for files and after choosing the

appropriate file; can load it to the system by using this section as seen in Figure 5:

Figure 5: Loading data from a data file

3.1.2 Capturing Online Data

This component has two parts as the following:

3.1.2.1 Web Crawler

This part enables an IT team member to crawl and download a web page that has useful experience data. By this engine, one can collect experience data from web pages, which he/she considers as useful. Crawled web pages are output of this sub-component.

Figure 6: WebSphinx user interface

After getting outputs of web Crawler, a user can use the Web Page Parser to create meaningful dataset files for the system.

Different web crawlers can be used to crawl useful web pages to collect data. One can use his/her web crawling software tool by embedding it into the Web Crawler component. WebSPHINX customizable web crawler [18] tool is used to crawl web pages. It is a well known and free web page crawler with its customizable parameters such as depth, thread number, page size etc. Figure 6 shows the form view of the WebSPHINX.

3.1.2.2 Web Page Parser

This component enables the proposed system to parse crawled data from web pages for the system data uploader. Web Crawler section outputs collection of folders with web page documents in them. By choosing the web crawl folders, which one can extract experience data from, and the output file, which one can save the extracted experience data, one will be able to parse the web pages from a collection of web documents into meaningful text files for the system. The interface for the Web Page Parser is shown in Figure 7.

Figure 7: Form window of the web page parser

After choosing the input folder and output document, the user need to choose one of the source pages from radio button group box. As mentioned before, crawling process goes as, firstly defining the useful web pages to crawl, then crawling process downloads these web pages to folders and from the corresponding documents. Every web page has its own html and text structure of its experience instance. Because of that, each web site needs its own parser. The software tool is an open source, so users can easily insert their own parsing function for a specific web site that they want to use. In this thesis, helpdesk web pages of Princeton University [19], Parallels [20] and Go Daddy [21] are used. To use the right parser for a web page, the user needs to choose the web page name from the radio button group box and this will call the convenient parser function for that the chosen web page. Finally, by clicking the "Parse the Web Source" button, preprocessing the downloaded and selected web source for experience dataset generates the text file for the system. Examples of output experience data text files are given in the following Data Sets subsection.

3.1.3 Datasets for the Data Base

In this thesis, three data sets are used to see the effectiveness of the proposed system. The data sets vary from each other according to the number of classes, the number of instances of a class, different area subjects, and different sources etc. All of the data sets consist of Frequently Asked Questions (FAQ) knowledgebases of commonly known Internet web sites as Princeton University IT Help Desk Web Site [19], which hereafter will be called as PrincetonDS, Parallels Free Support Resources Knowledgebase Web Site [20], which hereafter will be called as ParallelsDS and Go Daddy Help Center Web Site [21], which hereafter will be called as GoDaddyDS. In this thesis, 100 instances are selected randomly from each data set for evaluating the proposed system. Creating a dataset for a webpage is done by doing following steps: (i)Data is crawled from the selected subject of web page to a folder(Each category has its own folder to be sure that the data collected belongs to that category. For example PIHK(Princeton IT helpdesk knowledgebase web site) has 5 folders) (ii)Instances are labeled depending on their folders so depending on the category that they are collected

from(For example, instances in Navigating the Web folder of PIHK are labeled as 3) (iii)Predefined number of instances are collected randomly from the related folders(For example, 20 instances are collected randomly from the Navigating the Web folder of PIHK) (iv)Randomly collected instances are merged together (For example, randomly collected instances from folders of PIHK are merged together to create PrincetonDS dataset). Finally, datasets as PrincetonDS with 100 instances, ParallelsDS with 100 instances and GoDaddyDS with 100 instances are created.

Each instance of all datasets consist of two parts as "Title" and "Solution". These two parts of an instance are defined depending on the structure of the source's FAQ Knowledgebase. These are detailed in this section. After crawling the aforementioned web sites, preprocessing is performed on each data instance. Preprocessing involves stemming, stop word removal and punctuation removal of all data collected from the sources. Additionally, since different structures are used at each web site (such as for the definition of a problem sometimes title, or synopsis is used), to have one common name, "Title" is used to define the problem for a FAQ. Sources, preprocessing and datasets are explained in detail in the following sections.

3.1.3.1 PrincetonDS

The source of the PrincetonDS is Princeton University IT Help Desk knowledgebase [19]. Original knowledgebase consists of eleven categories and many subcategories under each one. Each subcategory consists of problems with their solutions. A short description of each problem is represented under the subcategory leading to a detailed description of the problem and the solution.

At this knowledgebase, FAQ (problem and solutions pairs under subcategories) has a structure, which defines problem and solution pairs under three titles. It defines the problem under "Title" section by the representative description which is also used under subcategories to guide the user. Additionally, a problem is defined under the "Synopsis" heading by the sentences which include key words for the problem. Finally, the solution is defined under the "Solution" heading. Solution section sometimes contains related

links for the defined problem but mostly it consists of textual information, which gives directly the solution for the problem. Figure 8 shows a sample of FAQ, which is taken from Princeton IT Help Desk Website, Desktop Computing main category, Antivirus Protection subcategory [19].

The five categories are used to choose records from the PrincetonDS dataset. These five main categories are “Desktop Computing”, “Services and Facilities”, “Navigating the Web”, “Operating Systems” and “Policies and Guidelines”. Twenty question-answer pairs are chosen from each of these categories. Deciding which pair to use in the dataset is done randomly and it is independent from the subcategories. “Desktop Computing” category is related to concepts such as antivirus information, questions about text processing tools, questions about spreadsheet processing tools, questions about database processing tools and questions about compressed files etc. “Services and Facilities” category is related to concepts such as getting help about software installations, hardware upgrades, buying a computer or equipment, data recovery, accounts and passwords etc. “Navigating the Web” category is related to concepts such as web browsers, e-mail applications and attachments etc. And “Operating Systems” category is related to concepts such as Windows, Macintosh and Unix operating systems. Finally the last category “Policies and guidelines” is related to concepts such as information security and computing guidelines.

After selecting twenty question-answer pairs from each of the categories, all question-answer pairs are put together into a text file. During this process, the structure of the original knowledgebase instance is converted into the instance structure that is defined as “Title and Solution”. “Title” and “Synopsis” section of the each instance is defined as the Title section of the dataset instance. The “Solution” section of the PrincetonDS instance is defined as the Solution section of the dataset instance. Figure 9 shows a sample of Title and Solution instance for my dataset.

After creating dataset text file, punctuation removal, stop word removal and stemming preprocesses are done to create PrincetonDS dataset input file. The Porter Stemming Algorithm is used in stemming step of preprocessing [23].

Title: Sasser Worm: How to download the removal tool and fix your computer

Synopsis:
Sasser Worm: How to fix and remove from your computer

Solution:
If your computer is infected with the Sasser worm, you may experience one or more of the following symptoms:

- Your computer performance is decreased or your network connection is slow.
- You may see a dialog box that contains text that refers to LSA Shell.
- Your computer may restart every few minutes without user input.

Removal Instructions

1. Windows XP (Home and Professional versions) must first turn off System Restore prior to running the removal tool.

- Right-click on **My Computer**
- Select **Properties**
- Click the **System Restore** tab
- Check the box **Turn off System Restore**
- Click **Apply**
- Click **OK**

2. Download the McAfee Stinger Removal Tool at: http://vil.nai.com/vil/stinger. Save it to your Desktop.

3. Disconnect from all network connections. This is a very important step; if you do not, you will just reinfect yourself immediately.

- Unplug your Ethernet connection from your computer
- Disable Wireless

4. Double-click on the Stinger Removal Tool to run the program.

5. Update security patches by using the Windows Update system. See: http://kb.princeton.edu/9501.

6. Update your Symantec virus definitions. See: http://kb.princeton.edu/8909.

7. Reboot your computer to complete the process.

Note for Windows 95/98/Me computers:
According to Symantec: "The W32.Sasser family of worms can run on (but not infect) Windows 95/98/Me computers. Although these operating systems cannot be infected, they can still be used to infect vulnerable systems that they are able to connect to. In this case, the worm will waste a lot of resources so that programs cannot run properly, including our removal tool. (On Windows 95/98/Me computers, the tool should be run in Safe mode.)"

Viruses are running rampant on campus and around the world. Learn about the top ten things you could be doing that is putting your computer and the intellectual property on your computer at risk. Please read **The Top Ten Reasons Your Computer Will Get Infected With A Virus and What You Should Do About It** at the URL: http://kb.princeton.edu/9544.

Figure 8: Sample of Princeton University IT Help Desk Web Site problem/solution [22]

Title:

Sasser Worm: How to download the removal tool and fix your computer
Sasser Worm: How to fix and remove from your computer

Solution:

If your computer is infected with the Sasser worm, you may experience one or more of the following symptoms:

Your computer performance is decreased or your network connection is slow.
You may see a dialog box that contains text that refers to LSA Shell.
Your computer may restart every few minutes without user input.

Removal Instructions

1. Windows XP (Home and Professional versions) must first turn off System Restore prior to running the removal tool.
   Right-click on My Computer
   Select Properties
   Click the System Restore tab
   Check the box Turn off System Restore
   Click Apply
   Click OK
2. Download the McAfee Stinger Removal Tool at: http://vil.nai.com/vil/stinger. Save it to your Desktop.
3. Disconnect from all network connections. This is a very important step; if you do not, you will just reinfect yourself immediately.
   Unplug your Ethernet connection from your computer
   Disable Wireless
4. Double-click on the Stinger Removal Tool to run the program.
5. Update security patches by using the Windows Update system. See: http://kb.princeton.edu/9501.
6. Update your Symantec virus definitions. See: http://kb.princeton.edu/8909.
7. Reboot your computer to complete the process.

Note for Windows 95/98/Me computers:According to Symantec: "The W32.Sasser family of worms can run on (but not infect) Windows 95/98/Me computers. Although these operating systems cannot be infected, they can still be used to infect vulnerable systems that they are able to connect to. In this case, the worm will waste a lot of resources so that programs cannot run properly, including our removal tool. (On Windows 95/98/Me computers, the tool should be run in Safe mode.)"

Viruses are running rampant on campus and around the world. Learn about the top ten things you could be doing that is putting your computer and the intellectual property on your computer at risk. Please read The Top Ten Reasons Your Computer Will Get Infected With A Virus and What You Should Do About It at the URL: http://kb.princeton.edu/9544.

Figure 9: Title and Solution instance sample

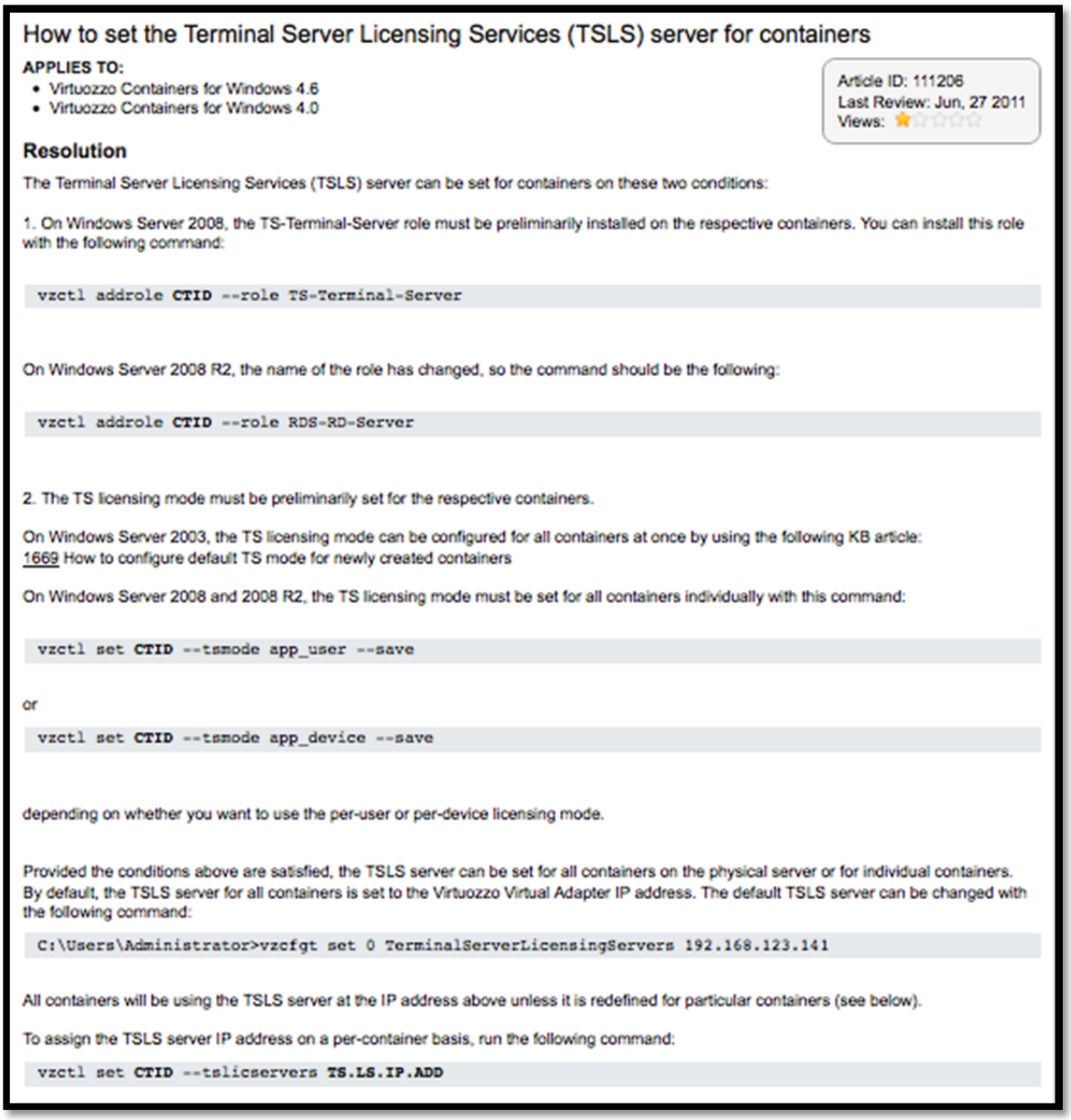
3.1.3.2 ParallelDS

The source of the ParallelsDS is Parallels Free Support Resources Knowledgebase Web Site [20]. Parallels is a virtualization and automation software that is used to optimize computing for consumers, businesses and service providers across all major hardware operating system, and virtualization platforms [24]. Original knowledgebase consists of ten titles to provide solutions for users under different subjects such as “Parallels Desktop for Mac”, “Parallels Desktop for Upgrading to Windows 7”, “Parallels Server”, “Parallels Automation” etc. Some main subjects guides you to subcategories under them and some main subjects guides you directly to FAQ which are named as articles. Articles consist of detailed description of the problem and the solution.

At this web site, articles has a structure which define problem and solution pairs sometimes under a title and two sections, sometimes under a title and three sections. At some articles, it defines the problem under a title and only “Applies To” section. At some articles, it defines the problem under a title and “Applies To” section with “Symptoms” section. Title of the article is a description of the problem and solution. “Applies To” section shows the costumer what this solution can be applied to and “Symptoms” section contains the question or definition for the problem. The article finally defines the solution under “Resolution” section. Resolution section commonly consists of textual information for solution sometimes with downloadable attachments and sometimes with demonstration pictures. Figure 10 shows a sample article which is taken from Parallels Free Support Resources Knowledgebase Web Site, Parallels Vituozzo Containers category, How to set the Terminal Server Licensing Services (TSLS) server for containers [20].

The eight main categories on this web site are used to choose question-answer pairs, which construct ParallelsDS dataset. The number eight is chosen randomly just to see how different it would be compared to the five chosen from the PrincetonDS web site, because it is wanted to see how the algorithms behave on datasets with instances gathered from different number of categories. These eight categoies are “Parallels Desktop for Mac”, “Parallels Desktop for Upgrading to Windows 7 Licensing, Activation

and Registration", "Parallels Sphera", "Parallels Helm", "Parallels Confix", "Parallels SiteStudio" (Sphera, Helm, Confix and SiteStudio categories are available under the "More Products" section of the main categories) , "Parallels Plesk Products" and



How to set the Terminal Server Licensing Services (TSLS) server for containers

APPLIES TO:

- Virtuozzo Containers for Windows 4.6
- Virtuozzo Containers for Windows 4.0

Article ID: 111206
Last Review: Jun, 27 2011
Views:

Resolution

The Terminal Server Licensing Services (TSLS) server can be set for containers on these two conditions:

1. On Windows Server 2008, the TS-Terminal-Server role must be preliminarily installed on the respective containers. You can install this role with the following command:

```
vzctl addrole CTID --role TS-Terminal-Server
```

On Windows Server 2008 R2, the name of the role has changed, so the command should be the following:

```
vzctl addrole CTID --role RDS-RD-Server
```

2. The TS licensing mode must be preliminarily set for the respective containers.

On Windows Server 2003, the TS licensing mode can be configured for all containers at once by using the following KB article:
1669 How to configure default TS mode for newly created containers

On Windows Server 2008 and 2008 R2, the TS licensing mode must be set for all containers individually with this command:

```
vzctl set CTID --tsmode app_user --save
```

or

```
vzctl set CTID --tsmode app_device --save
```

depending on whether you want to use the per-user or per-device licensing mode.

Provided the conditions above are satisfied, the TSLS server can be set for all containers on the physical server or for individual containers. By default, the TSLS server for all containers is set to the Virtuozzo Virtual Adapter IP address. The default TSLS server can be changed with the following command:

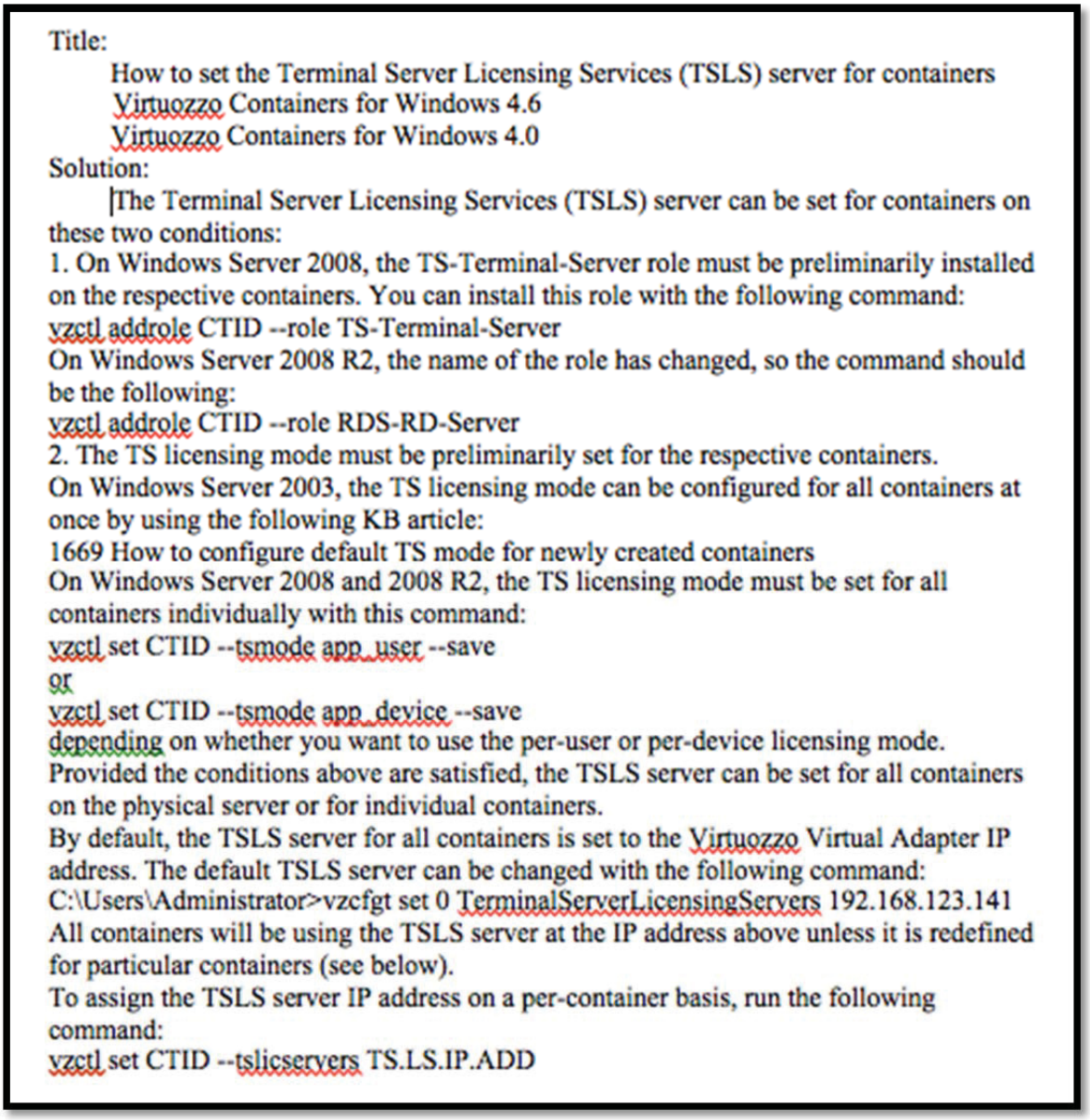
```
C:\Users\Administrator>vzcfgt set 0 TerminalServerLicensingServers 192.168.123.141
```

All containers will be using the TSLS server at the IP address above unless it is redefined for particular containers (see below).

To assign the TSLS server IP address on a per-container basis, run the following command:

```
vzctl set CTID --tslicservers TS.LS.IP.ADD
```

Figure 10: A sample of Parallels Free Support Resources Knowledgebase Web Site article [25]



Title:

How to set the Terminal Server Licensing Services (TSLS) server for containers Virtuozzo Containers for Windows 4.6 Virtuozzo Containers for Windows 4.0

Solution:

The Terminal Server Licensing Services (TSLS) server can be set for containers on these two conditions:

1. On Windows Server 2008, the TS-Terminal-Server role must be preliminarily installed on the respective containers. You can install this role with the following command:

vzctl addrole CTID --role TS-Terminal-Server

On Windows Server 2008 R2, the name of the role has changed, so the command should be the following:

vzctl addrole CTID --role RDS-RD-Server

2. The TS licensing mode must be preliminarily set for the respective containers.

On Windows Server 2003, the TS licensing mode can be configured for all containers at once by using the following KB article:

1669 How to configure default TS mode for newly created containers

On Windows Server 2008 and 2008 R2, the TS licensing mode must be set for all containers individually with this command:

vzctl set CTID --tsmode app_user --save

or

vzctl set CTID --tsmode app_device --save

depending on whether you want to use the per-user or per-device licensing mode.

Provided the conditions above are satisfied, the TSLS server can be set for all containers on the physical server or for individual containers.

By default, the TSLS server for all containers is set to the Virtuozzo Virtual Adapter IP address. The default TSLS server can be changed with the following command:

C:\Users\Administrator>vzcfgt set 0 TerminalServerLicensingServers 192.168.123.141

All containers will be using the TSLS server at the IP address above unless it is redefined for particular containers (see below).

To assign the TSLS server IP address on a per-container basis, run the following command:

vzctl set CTID --tslicservers TS.LS.IP.ADD

Figure 11: An article sample for the dataset

"Parallels Virtuozzo Containers". Respectively, twenty six, twenty, fifteen, twelve, eight, seven, seven and five pairs are chosen randomly from these categories. The number of question-answer pairs are chosen randomly.

After defining articles from each of the defined eight categories, all articles put together into a text file. During this process, the structure of the original instance is

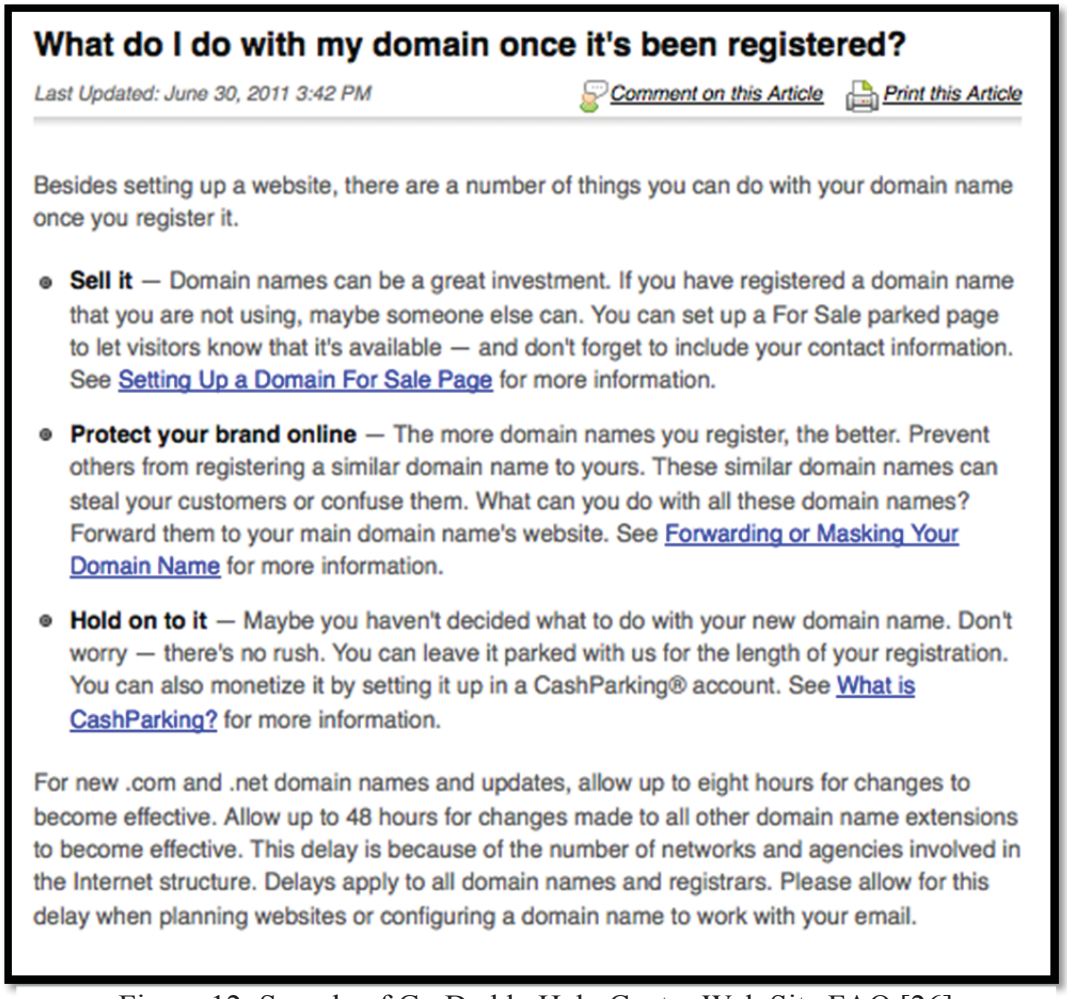
converted into the instance structure that is defined as “Title and Solution”. The title, “Applied To” and “Symptoms” (if exists) sections of the each instance is defined as the Title section of the dataset instance. “Resolution” section of the instance is defined as the Solution section of the dataset instance. Figure 11 shows a sample of Title and Solution instance for the dataset. After creating the dataset text file, punctuation removal, stop word removal and stemming preprocesses are done to create ParallelsDS dataset input file. The Porter Stemming Algorithm is used in stemming step of preprocessing [23].

3.1.3.3 GoDaddyDS

The source of the GoDaddyDS is Go Daddy Help Center Web Site [21]. Original web site consists of nine titles to provide solutions for users under different subjects of IT such as “Domains”, “Hosting & Servers”, “Site Builders”, “Security” etc. Each main subject guides you to subcategories under them. Each subcategory consists of question-answer pairs as FAQ or articles that contain a problem with its solutions under that subcategory. Short description of each problem is represented under the subcategory to lead you to the detailed description of the problem and the solution.

At this web site, FAQs and articles have a structure which defines problem and solution pairs under a title and a solution text. It defines the problem in the title and the solution is defined after the title. Solution sometimes contains related links for the defined problem. Figure 12 shows a sample of FAQ which is taken from Go Daddy Help Center Web Site, “Domains” subject, “Registering Domain Names” subcategory [21].

The three main categories at the web site are used to choose question-answer pairs, which construct GoDaddyDS dataset. The number three is chosen randomly just different than five and eight condition, because it is wanted to see how the algorithms perform on datasets with instances gathered from different number of categories. These three categories are “Registering Domain Names” under Domains section, “SSL Certificates” under Security section, “Express Email Marketing” under Business section. Respectively, forty one, thirty four and twenty five pairs are chosen randomly from the aforementioned categories.



What do I do with my domain once it's been registered?

Last Updated: June 30, 2011 3:42 PM *Comment on this Article* *Print this Article*

Besides setting up a website, there are a number of things you can do with your domain name once you register it.

- **Sell it** — Domain names can be a great investment. If you have registered a domain name that you are not using, maybe someone else can. You can set up a For Sale parked page to let visitors know that it's available — and don't forget to include your contact information. See Setting Up a Domain For Sale Page for more information.
- **Protect your brand online** — The more domain names you register, the better. Prevent others from registering a similar domain name to yours. These similar domain names can steal your customers or confuse them. What can you do with all these domain names? Forward them to your main domain name's website. See Forwarding or Masking Your Domain Name for more information.
- **Hold on to it** — Maybe you haven't decided what to do with your new domain name. Don't worry — there's no rush. You can leave it parked with us for the length of your registration. You can also monetize it by setting it up in a CashParking® account. See What is CashParking? for more information.

For new .com and .net domain names and updates, allow up to eight hours for changes to become effective. Allow up to 48 hours for changes made to all other domain name extensions to become effective. This delay is because of the number of networks and agencies involved in the Internet structure. Delays apply to all domain names and registrars. Please allow for this delay when planning websites or configuring a domain name to work with your email.

Figure 12: Sample of Go Daddy Help Center Web Site FAQ [26]

After downloading articles from each of the selected categories, all of them are put together into a text file. During this process, the structure of the original knowledgebase instance is converted into the instance structure that is defined as “Title and Solution” before. Title of the each original instance is defined as the Title part of the dataset instance. The solution part of the original instance is defined as the Solution section of the dataset instance. Figure 13 shows a sample of Title and Solution instance for the data set.

After creating dataset text file, punctuation removal, stop word removal and stemming preprocesses are done to create GoDaddyDS dataset input file. The Porter Stemming Algorithm is used in stemming step of preprocessing [23].

Title:

What do I do with my domain once it's been registered?

Solution:

Besides setting up a website, there are a number of things you can do with your domain name once you register it.

Sell it — Domain names can be a great investment. If you have registered a domain name that you are not using, maybe someone else can. You can set up a For Sale parked page to let visitors know that it's available — and don't forget to include your contact information. See Setting Up a Domain For Sale Page for more information.

Protect your brand online — The more domain names you register, the better. Prevent others from registering a similar domain name to yours. These similar domain names can steal your customers or confuse them. What can you do with all these domain names? Forward them to your main domain name's website. See Forwarding or Masking Your Domain Name for more information.

Hold on to it — Maybe you haven't decided what to do with your new domain name. Don't worry — there's no rush. You can leave it parked with us for the length of your registration. You can also monetize it by setting it up in a CashParking® account. See What is CashParking? for more information.

For new .com and .net domain names and updates, allow up to eight hours for changes to become effective. Allow up to 48 hours for changes made to all other domain name extensions to become effective. This delay is because of the number of networks and agencies involved in the Internet structure. Delays apply to all domain names and registrars. Please allow for this delay when planning websites or configuring a domain name to work with your email.

Figure 13: FAQ sample for the dataset

3.2 Text Clustering Engine

Text Clustering Engine is managed by the System Update module of the UI. This module starts the text clustering engine and at the end of the process it lets user to update the knowledgebase of the system depending on his/her preferences. The System Update module and its relation with the system is as show in Figure 14.

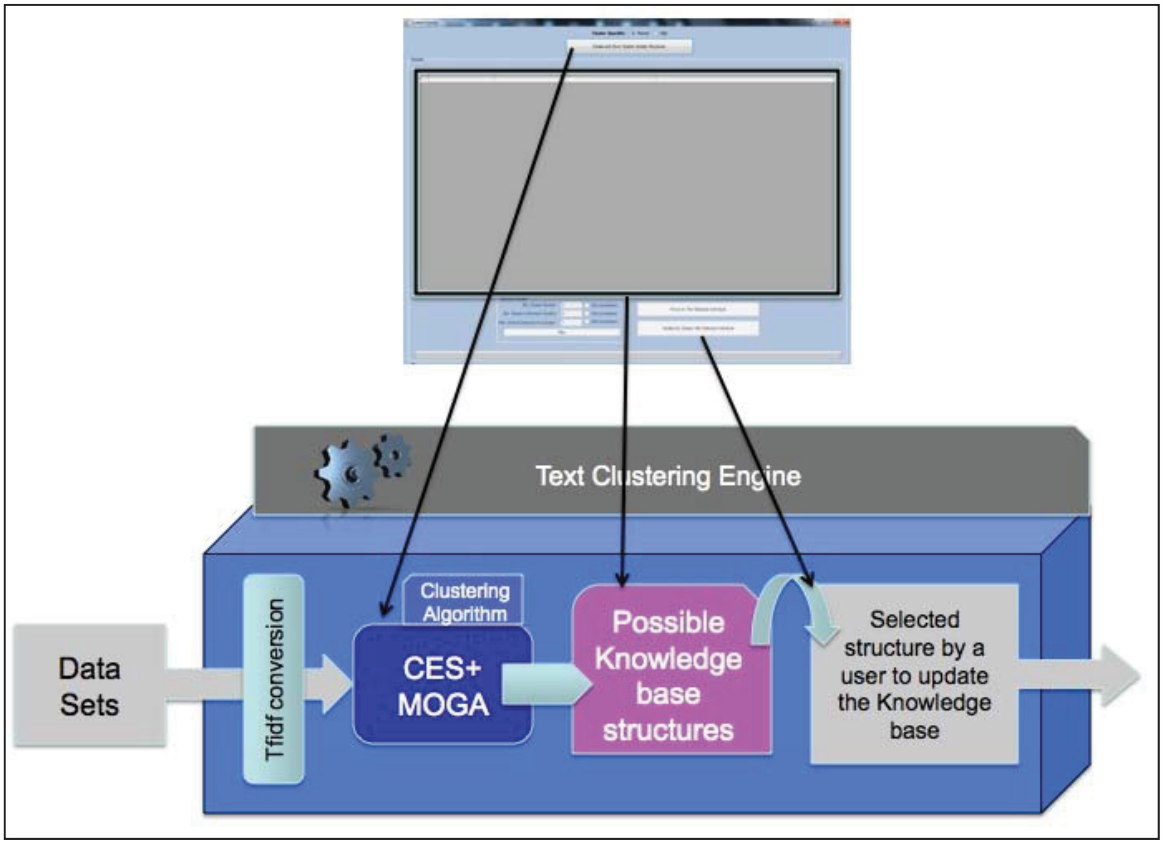


Figure 14: The System Update module

To select an efficient clustering algorithm different machine learning algorithms and distance measures are compared for the Text Clustering Engine. An evolutionary approach is used to automate the selected CES+ with the Multi Objective Genetic Algorithm (MOGA) and this is explained in following subsections.

Text clustering engine firstly converts text data into *tfidf* vectors for the clustering algorithm. It then shows the possible knowledgebase structures to the user. The *tfidf*

(term frequency – inverse document frequency) is a measure to define how important a word is for a document in a collection. With a simple description, the *tf* (term frequency) represents the appearance of a given term in a document. The *idf* (inverse document frequency) measures how rare the term is in the document collection and is calculated by dividing the total number of documents by the number of documents that contain the given term and then taking the logarithm of that division.

3.2.1 Employed Algorithms

This subsection introduces the basic concepts employed in this thesis, which are used to provide the best clustering output for users. Explaining the employed clustering methods namely Expectation Maximization Algorithm (EM), DBScan, K-Means algorithms and distance measuring methods, namely Cosine Similarity, Jaccard Index is a necessity to have a better understanding of the concept. This chapter will describe the machine learning techniques and distance measures evaluated for the proposed system. This is not a comprehensive explanation of these concepts, for more detailed information reader should refer to the cited material in each section.

3.2.1.1 Techniques for Clustering

3.2.1.1.1 Expectation Maximization (EM) Algorithm

The EM algorithm is an iterative optimization technique to estimate the maximum likelihood computation in the presence of hidden or missing data. In maximum likelihood computation, the most likely model parameters are estimated for the observed data.

EM algorithm's iterations have two processes: the expectation step and the maximization step. In the expectation step, the conditional expectation is used to estimate the missing data and the current estimate of the model parameters while the observed data is given [27]. In the maximization step, the missing data is assumed to be known and the

likelihood function is maximized with that assumption. Instead of the actual missing data, the estimate of the missing data from the expectation step is used. Briefly, at each iteration, the algorithm tries to increase the likelihood by constructing a local lower bound to the next distribution in expectation step and improving the estimate for the unknowns by optimizing the bound in the maximization step.

If it is assumed that data *A* is generated by some distribution, it can be also assumed a joint density function as in Eq.1:

$$p(c|\theta) = p(a,b|\theta) = p(b|a,\theta)p(x|\theta)$$

Eq. 1

where *A* is called the incomplete data and it is assumed a complete data set exists C=(A,B) [28].

The joint density usually comes up from the assumption of hidden variables and the parameter value guesses and the marginal density function p(a| $\boldsymbol{\theta}$). In other situations, a joint relationship between the missing and the observed values are assumed.

The first step of the algorithm finds the expected value of the complete data log likelihood p(X,Y| $\boldsymbol{\theta}$), it is called as E-step of the algorithm and is defined as in Eq.2:

$$Q\left(\theta,\theta^{i-1}\right) = E\left[log\ p(A,B|\theta)|A,\theta^{(i-1)}\right]$$

Eq. 2

where *B* is the unknown data, *A* is the observed data and $\boldsymbol{\theta}$ is the new parameters that is optimized to increase *Q*likelihood, $\boldsymbol{\theta}^{(i-1)}$ is the current parameters that is used to evaluate the expectation.

The second step of the algorithm tries to maximize the expectation that is computed in the E-step, and it is called as M-step. This step is represented as in Eq.3:

$$\theta^i = {}^{argmax}_{\theta} Q(\theta,\theta^{(i-1)})$$

Eq.3

The E-step and the M-step are repeated as much as necessary. Each iteration aims to increase the log-likelihood and the algorithm aims to converge to the local maximum of the likelihood function [29].

3.2.1.1.2 DBScan Algorithm

The DBScan algorithm is a density-based clustering algorithm, which is introduced by Martin Ester et al. in 1996 [30]. The algorithm is based on the center-based approach in which by counting the number of points within a specified radius, the density is estimated for a particular point in the data set [31]. The density of any point depends on the specified radius. If radius is too small then all the points will have the density of one and if the radius is too big then the all points will have the density of the same size with the instances. Figure 15 shows an illustration of the center based density approach:

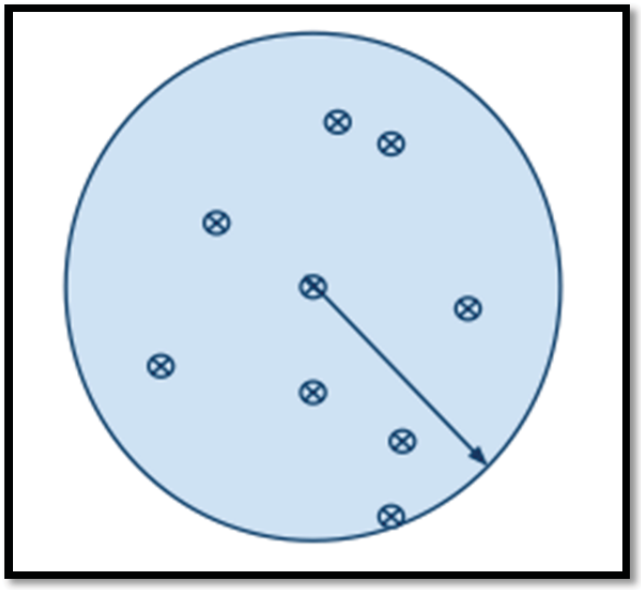


Figure 15: Center-based density of points

In the center-based approach, points can be classified as being a core point, which is in the interior of a dense region, a border point, which is on the edge of a dense region and a noise or background point, which is in a sparsely occupied region [20].

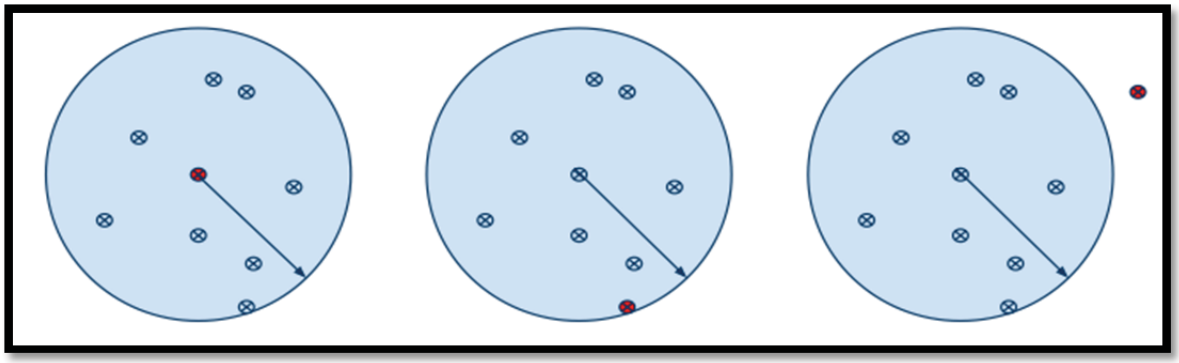


Figure 16: Core point, border point and noise point (right to left)

Core points are in the interior of the density based cluster. If the number of points in a defined neighborhood by the specified distance parameter and the distance function is more then the threshold then the point is called as a core point. Border points fall within the neighborhood of a core point and they are not core points. Any point that is not a border or a core point is called a noise point.

In Figure 16, a core point is seen, a border point and a noise point from left to right. The representative points are colored as red depending on their type. The arrow from core point to the border is the specified distance parameter for the radius, which is named as *Eps*.

The DBScan algorithm puts two core points that are close within a distance into the same cluster. Any other border point, which is close enough to a core point is put in the same cluster as the core point. Noise points are discarded. DBScan algorithm can be summarized as the following [20]:

- Define and label all points as core, border and noise
- Discard and eliminate noise points
- Put an edge between all core points that are in the distance of *Eps* of each other
- Create a separate cluster by using each group of connected core points
- Assign each border point to one of the clusters of its associated core points

DBScan is an effective algorithm for data that has noise and it. It can handle clusters of arbitrary shapes and sizes. So it can find many clusters that for example K-Means clustering algorithm can not find because of the cluster shape. But DBScan may have difficulty with high dimensional data because of difficulty in defining the density. Additionally, it can be expensive if the data is high dimensional because of computing all pair wise proximities of nearest neighbors.

3.2.1.1.3 K-Means Algorithm

One of the most commonly used clustering algorithms is K-Means algorithm [28,33], which is declared as a nonhierarchical and partitional clustering technique [34].

In K-Means clustering, number of clusters, which is represented by "*k*", are assumed to be fixed. New clusters are created by using the distance defined from the center of each cluster. This is also considered as the error quantity or the quality of the clustering. The quality of the clustering is determined by the error function as following as in Eq.4:

$$E = \sum_{y=1}^{k} \sum_{xl \in Cy} \left| x_l - w_y \right|^2$$

Eq.4

where *k* amount of samples ($w_1, \dots, w_k$) are initialized to one of the input patterns ($x_1, \dots, x_n$) so

$$w_y = x_l, y \in \{1, \dots, k\}, l \in \{1, \dots, n\}$$

To find suitable value of *k* for effective results, generally several values of *k* are employed. The number of iterations necessary can vary from thousands to small amounts. Number of clusters, number of patterns and the input data distribution can effect the number of iterations for effective results [35].

The Algorithm can be summarized as the following:

- Initialize k prototypes$(w_1, \ldots, w_k)$ where $w_y = x_l$, $y \in \{1, \ldots, k\}$, $l \in \{1, \ldots, n\}$
- Associate each cluster C_i with prototype w_j
- Repeat until cluster membership no longer changes or no significant change in E

for each input vector x_l , where $l \in \{1, \ldots, n\}$,

assign x_l to the cluster C_{j*} with nearest prototype w_{y*}

for each cluster C_y, where $y \in \{1, \ldots, k\}$

update the prototype w_y to be the centroid of all samples

currently in C_y, so $w_y = \sum_{xl \in C_y}^{k} x_l / |C_y|$

Compute $E = \sum_{y=1}^{k} \sum_{xi \in C_y} |C_y|$

A sample space can be considered as following for the algorithm as in Figure 17.

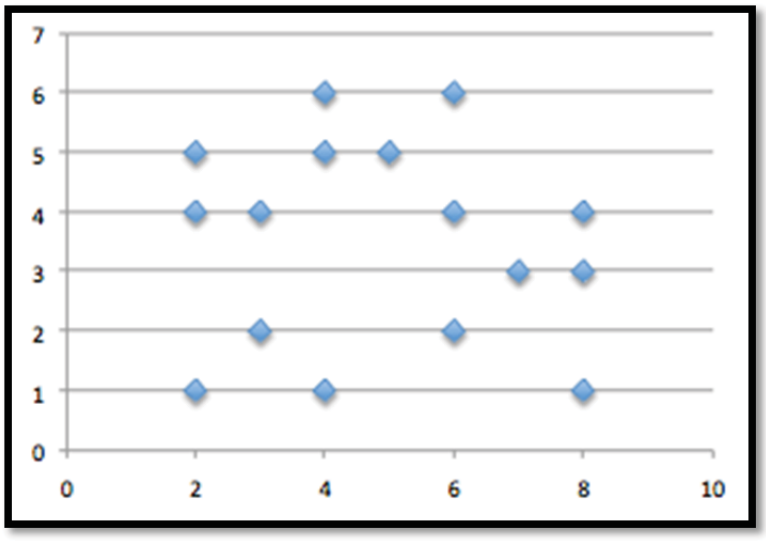


Figure 17: Instances on a sample space

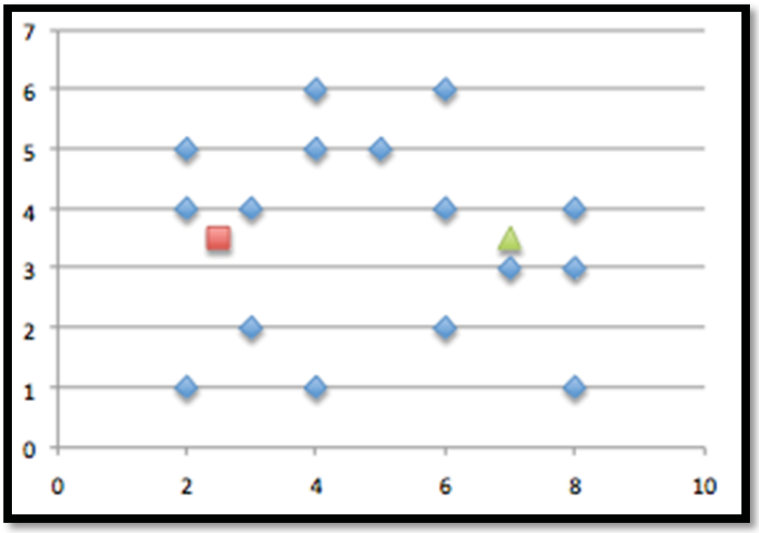


Figure 18: Two cluster centers are defined

Two centers of clusters are defined as points on the sample space as in Figure 18. These center points are going to be used as the representatives of the cluster and then they are going to be used for defining cluster elements.

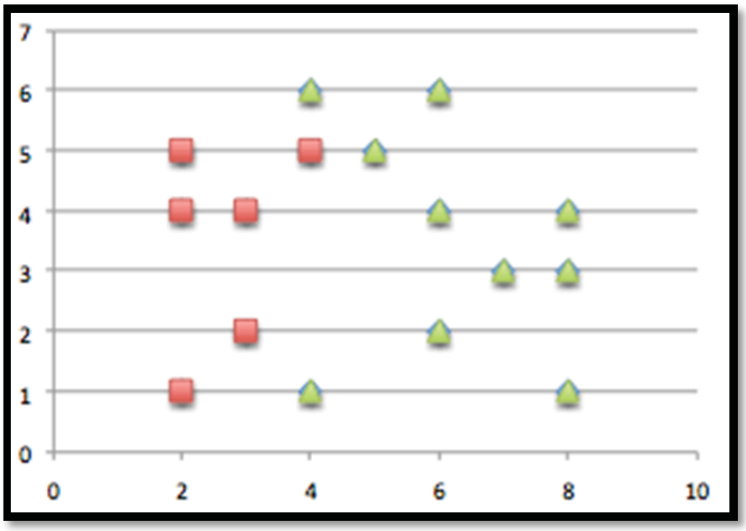


Figure 19: Instances of two clusters are defined

All instances are clustered depending on their distances to the cluster centers, for example by using Euclidean or Manhattan distance as in Figure 19.

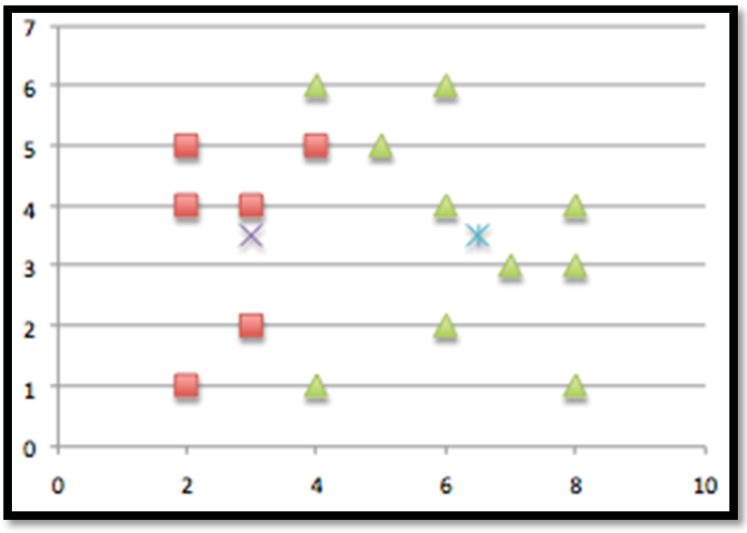


Figure 20: New defined two cluster centers

New centers for clusters are defined or the centers of the clusters are moved as in Figure 20.

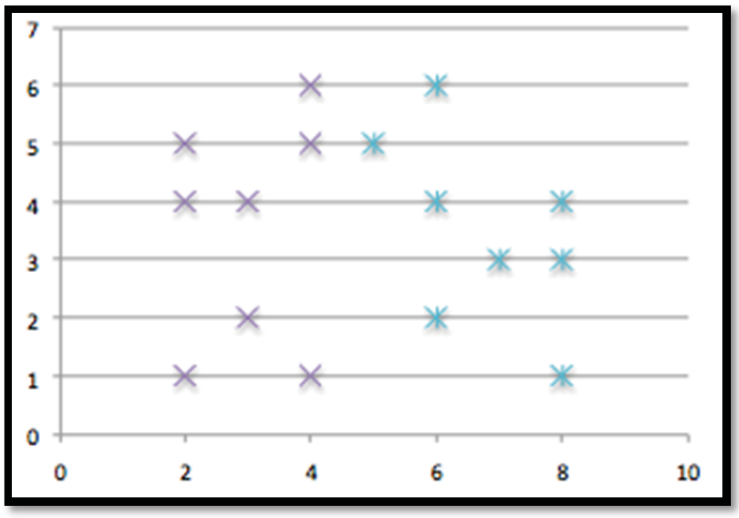


Figure 21: New defined two clusters

After moving the centers, it is possible to have some instances closer to new centers. So clusters are updated depending on the new distances of the instances to the new cluster centers as in Figure 21. Finally the new clusters are received after the new iteration.

3.2.1.1.4 CES

The CES Algorithm is a clustering algorithm that is created to have a high efficiency in text and document clustering. It has three main steps as creating core clusters by high threshold value, expanding clusters by low threshold value and sophistication of the clusters.

In the creating core clusters step, small clusters are created from the threads. Then the dictionary is constructed. The core clusters need to be constructed precisely by using similar threads. The similarity between threads are calculated by addition of the similarities of questions and the similarities of the answers as in Eq.5:

$$Similarity(Document_1, Document_2) = (1-\propto) \times CosSimQ + \propto \times CosSimA$$

Eq.5

where α is the content value, which specifies the rate that defines the effect of question and answer similarity result on total similarity, and where CosSim() represents Cosine similarity measure between two vectors as in Eq.6:

$$CosSimQ = \frac{\vec{Q}_1 . \vec{Q}_2}{|\vec{Q}_1| \times |\vec{Q}_2|} , CosSimA = \frac{\vec{A}_1 . \vec{A}_2}{|\vec{A}_1| \times |\vec{A}_2|}$$

Eq.6

During the process of this step, *tf-idf* of the all words in the vectors are calculated and stored as a feature vector of the threads for the words in the thread and a feature vector of the clusters for the words in the cluster. The dictionary of the clusters consist of *tf-idf* value of the words in each cluster. The words, whose *tf-idf* value exceeds the given threshold value for creating the dictionary, are added to the dictionary of the cluster. So that means the words, which are important for the definition of the cluster behavior are

contained by the dictionary. In all steps of the CES algorithm, the dictionary is updated whenever the clusters are updated.

In the Expansion of the clusters step, the core clusters, which are generated in the first step are expanded by combining different clusters or adding a thread to a cluster. Because of the high threshold value in the creating core clusters step, it is possible to have many small clusters. The aim of this step is merging similar question and answer pairs. During the expanding process, the dictionary of the clusters are updated.

To expand clusters by merging them, the feature vector of the clusters is used to calculate similarity between clusters. The similarity between two clusters is measured as in Eq.7:

$$Similarity(Cluster_1, Cluster_2) = \overrightarrow{tfidf_1}.\overrightarrow{tfidf_2}$$

Eq.7

$$\frac{\overrightarrow{tfidf_{Q_1}}.\overrightarrow{tfidf_{Q_2}} + \overrightarrow{tfidf_{A_1}}.\overrightarrow{tfidf_{A_2}}}{2}$$

where $\vec{tfidf}$ represents the feature vector of the cluster. Using feature vectors of the clusters, lets the similarity result be retrieved from the words, which are more related to the characteristics of the cluster.

To expand clusters by adding new threads to them, the feature vector of the clusters and the word vector of the threads are used. The word vector of the thread contains the frequency of each word occurrence in that thread. The similarity between a cluster and a thread is measured by Cosine similarity between the thread's word vector and the cluster's feature vector. If the result value exceeds the defined threshold value, thread is added to that cluster. The similarity between a cluster and a thread is defined as in Eq.8:

$$Similarity(Cluster_C, Document_D) = (1-\propto) \times CosSimQ + \propto \times CosSimA$$

Eq.8

where

$$CosSimQ = \frac{\sum_{i=0}^{n} CosSimQ_i}{n}\ , CosSimQ = \frac{\sum_{i=0}^{n} CosSimA_i}{n},$$

and where

$$CosSimQ = \frac{\overrightarrow{tfidf_{Qc}} \,.\, \vec{Q}_D}{\left|\overrightarrow{tfidf_{Qc}}\right| \times \left|\vec{Q}_D\right|}\ , CosSimQ = \frac{\overrightarrow{tfidf_{Ac}} \,.\, \vec{A}_D}{\left|\overrightarrow{tfidf_{Ac}}\right| \times \left|\vec{A}_D\right|}$$

α is the content value, n is the number of threads in the $Cluster_C$, *tfidf* is the feature vector of the $Cluster_C$.

In the Sophistication of the Clusters step, the threads, which belong to another content, are extracted from a cluster. In the Expansion of the clusters step, threshold value is low so this can cause adding wrong threads to wrong clusters. The extraction of the thread is decided by *C(i)*. This index shows the number of words, which are characterizing the cluster, contained by the thread. *C(i)* is declared as in Eq.9:

$$C(i) = \sum_{Wi \in Dictionary(Cluster_l)} f(Wi)$$

Eq.9

where $Dictionary(Cluster_C)$ is the words of $Cluster_C$ in the dictionary

$$f(w_i) = \frac{Word(w_i)(tfidf)}{Number\ of\ words}$$

where $Word(w_i)(tfidf)$ means *tfidf* value of $w_i \in Dictionary(Cluster_C)$ of the corresponding thread. If the *C(i)* is lower than the specified threshold then the thread is extracted.

The CES algorithm is a clustering technique that employs three steps. However, because of comparisons and the number of updates employed during the clustering process, it has a high computational cost compared to many other commonly used clustering algorithms.

3.2.1.2 Techniques for Similarity Calculations

3.2.1.2.1 Cosine Similarity Measure

In Cosine similarity, similarity of two vectors is measured by the cosine of the angle between them [34]. If the angle between two vectors is zero then the result of the cosine function is one. The Cosine of the angle between two vectors determines if they are pointing to the same direction. For document clustering, usually the attribute vectors are the term frequency vectors of the documents. If two vectors as $\vec{A}$ and $\vec{B}$ are

considered, their similarity can be defined by using Cosine measure as in Eq.10:

$$sim(\vec{X}, \vec{Y}) = \frac{\sum_{i=1}^{n} X_i x Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2}\, x \sqrt{\sum_{i=1}^{n} (Y_i)^2}}$$

Eq.10

Similarity result +1 means the two vectors are exactly the same and -1 means two vectors are exactly opposite. Result 0 means the two vectors are independent and middle values represents the intermediate similarity between two vectors. In information retrieval process, Cosine similarity value ranges from 0 to 1 because term frequencies of the documents can not be negative.

3.2.1.2.2 Jaccard index

The Jaccard index is also known as the Jaccard Similarity Coefficient [36]. It is a measurement, which is used for identifying the similarity between two data vectors. In the simplest way, the Jaccard is defined index by defining two data sets as X and Y. The intersection region of these two sets is defined as $(X \cap Y)$ and the union region of these two sets are defined as $(X \cup Y)$. The Jaccard index is calculated as in Eq.11 based on these definitions:

$$\text{Jaccard}(X, Y) = \frac{X \cap Y}{X \cup Y} \qquad \text{Eq.11}$$

Assume two vectors as $\vec{A}$ and $\vec{B}$, the intersection and the union of these two vectors are required to calculate the Jaccard Distance between them. The inter- section between these two vectors can be defined by calculating the cardinality of the same value attributes as the following:

$$\vec{A} \cap \vec{B} = \sum_{i=0}^{n} A_i \cap B_i \quad , \quad A_i \in \vec{A} \quad \text{and} \quad B_i \in \vec{B}$$

where

$$A_i \cap B_i = \begin{Bmatrix} 1, & if\ A_i = B_i \\ 0, & otherwise \end{Bmatrix}$$

The union of the vectors $\vec{A}$ and $\vec{B}$ can be defined as the following:

$$\vec{A} \cup \vec{B} = \sum_{i=1}^{n} A_i \cup B_i \ , \ \ A_i \in \vec{A} \ ; B_i \in \vec{B} \quad \text{and} \quad A_i \cup B_i = 1$$

So the Jaccard index or Jaccard similarity measurement between two vectors is as in Eq.12:

$$Jaccard(\vec{A}, \vec{B}) = \frac{cardinality\ of\ intersection}{cardinality\ of\ union}$$

Eq.12

3.2.1.3 Genetic Algorithms

In the 1960s and the 1970s, John Holland invented Genetic Algorithms (GAs) in order to mimic the mechanisms of natural adaptation for computer systems [37]. Many computational problems such as prediction in financial markets, protein engineering etc. require adaptive computer programs for changing conditions. If a problem has a big solution space, then GAs are useful for the task of searching that space.

GAs have basic elements as chromosomes, fitness functions, crossover operators and mutation operators. Chromosome populations are used to represent candidate solutions in a search space. They consist of strings of bits or values, which are named as genes. In short, these are encoded solutions for the problem at hand. Selection of individuals (solutions) for new generations and to define parents is one of the most important concepts in evolution. Survival of the fittest in the changing environment is the way to create better generations. The fitness function is the element of GAs, which is used for selection of the fittest in the population to improve the individuals over generations. The aim of the fitness function is to define a set of parameter values that maximizes the multi-parameter function. Choosing fittest individuals for new generations is vital for reaching a better solution. Therefore defining a fitness function that is able to discriminate between fitter and other individuals is very important for the solution space. Crossover and Mutation are operators used for creating new offsprings from parents. The crossover operator chooses a certain position in the string of two chromosomes (parents) and exchanges materials of chromosomes depending on this certain point to generate a new offspring. Additionally, the mutation operator is used to flip some of the genes in the chromosomes of a new offspring to create diversity in the population. A simple genetic

algorithm can be outlined as the following:

1 Initialize the population with randomly generated n number of candidate solutions

2 Evaluate the fitness of each individual

3 Repeat for the defined number of generations G

- Select two best individuals from the current population depending on their fitness

- Crossover two selected parents

- Mutate the new individuals and insert them to the population

4 Replace the worst two individuals in the current population with the newly generated offsprings and go to step 2

An iteration of this algorithm is called as a generation and a set of generations is called a run. The number of iterations in a run can differ depending on the initial population, the fitness function and the operators employed.

Multi-Objective Genetic Algorithms

Optimization aims to find the best solution for a given problem given a set of objectives [38]. If an optimization problem involves only one objective, the searching for an optimal solution in a space is called as single-objective optimization [39]. Evolutionary algorithms are examples of such algorithms. However, in many problem solving tasks, usually there is more than one objective that is wanted to be optimized. If an optimization problem consists of at least two objectives then the searching for an optimal solution in the solution space is called as multi-objective optimization. In real world, most of the optimization problems involve multiple objectives. Sometimes, it is possible to have these objectives in conflict with each other. For example, when buying a plane ticket, it is wanted to minimize its cost while maximizing its comfort. Trying to optimize more than one objective for a defined problem always gives a set of solutions, which is called as Pareto Optimal Set. Individuals in Pareto Optimal Set are called as

nondominated and in the plot of objective functions, these nondominated individuals in Pareto Optimal Set are called as Pareto front. Goldberg suggested to assign ranks for the individuals in a population to define their domination [39]. The aim of giving ranks is to discriminate the Pareto nondominated set, which has the highest ranks, from the rest of the population. In the literature, several Multi Objective Evolutionary Algorithms are developed and the most well-known ones are :

1) Multi Objective Genetic algorithm (MOGA): The ranking used to define the best individuals. The rank of an individual indicates the number of other individuals, which it is dominated by. Nondominated individuals are assigned the possible highest fitness value. On the other hand, dominated individuals share the fitness of a region, which they belong. For more details, you can refer to [40].

2) Nondominated Sorting Genetic Algorithm (NSGA): The ranking is based on the nondomination and all nondominated individuals are classified as one category. The fitness value is proportional to the population size and to keep diversity in the population, individuals of a category share their fitness values. After that these individuals are ignored and then another layer of nondominated individuals are considered. Until getting all individuals in the population classified, this process continues. For more details, you can refer to [41].

3) Niched Pareto Genetic Algorithm (NPGA): Two individuals are randomly selected for a competition and the winner of the competition is put into the next generation. For the competition, a set of individuals are randomly chosen from the existing population. The competitor individual, which is nondominated by the individual set, wins the competition. If there is another situation, the result of the competition is decided by fitness sharing. This cycle keeps going on until the next generation is defined. For more details you can refer to [42].

3.2.2 CES+ and MOGA for Automation

3.2.2.1 CES+

The CES+ clustering algorithm is the improved version of the previous CES algorithm. In the CES+ algorithms, still the same defined formulas and equations are employed as well as the same steps (creating core clusters, expanding the clusters and sophisticating the clusters) as the CES algorithm. However, the difference between CES and CES+ algorithms comes from the implementation details by which the computational cost is minimized for time efficiency reasons. Therefore, some changes are made to the CES, which are derived from the proposed algorithm in [4].

The CES algorithm uses the same number of elements for all the vectors used in the algorithm such as the word vector, the feature vector etc. This is done by defining all words in all text documents as features (attributes) for all threads used in the algorithm. Indeed, this causes many attributes to exist in an instance even if many of them are not related to that instance. Unrelated attributes have zero value in an instance so these vectors end up being sparse with many zeros. Searching these long vectors for all comparisons and for all updates of clusters, results in a big computational cost for the original CES algorithm. So changing the vector structure is the first improvement made to reduce the computational cost. The structure of CES+, only uses the words, which are related to the document, as an attribute for the each document. So each word vector or feature vector of the document thread consists of the attributes, which are the words that exist in that document. So this gives a smaller vectors and non-sparse data. Cosine similarity is the basic measure that is used in the CES algorithm for comparing vectors. As it is seen in the Cosine similarity formula in Eq.13, zero values have no affect on the result of the formula if the multiplication is considered as zero when one of the attributes of a vector is not existing in the compared vector.

$$Cosine(\vec{A},\vec{B}) = \frac{\vec{A}.\vec{B}}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Eq.13

To create this structure, hash tables are used and it created great reduction in terms of computational cost.

The second improvement is that during the computation of similarity between two vectors. In this case, the vector, which has less attributes, is chosen as the base vector. Then the words of the base vector are searched to find the similarity, if it exists in the second vector, it is calculated. The attributes of the longer vector, which do not exist in the base vector, are not considered because they would be multiplied with zero and it has no affect on the summation section of the cosine similarity function Eq.13. So this again reduces the computational cost in terms of time.

3.2.2.2 Genetic Algorithm for Parameter Optimization

In this research, MOGA is employed to optimize the automation of CES+ algorithm from one data set to another. In this case, MOGA starts with an initial population of individuals, which are possible solutions for the problem, and evolves that initial population into better individuals, depending on the fitness function. Fitness function represents the objectives that are wanted to achieve. This work followed the MOGA framework proposed by Kim et al. [43] and Bacquet [44], but modifies the evolutionary component of Pareto Converging Genetic Algorithm. The algorithm converges towards the Pareto front, which consist of non-dominated individuals (solutions). Instead of replacing the entire population after each generation, two individuals are combined and the resulting two offsprings replace the two worst performing individuals in the population after each generation.

In GAs, two important concepts are representation of an individual and the fitness function. The representation of an individual defines how the member of the population represents solutions for the problem that is defined. The fitness function defines how the best individuals are chosen to create better generations. The representation of an individual is explained in section 2.2.1 and the fitness function is explained in section 3.2.2.2.2.

3.2.2.2.1 Individual Representation

Each individual in the population represents a set of CES+ algorithm's parameter values. As mentioned in section 4.2 and 4.3, CES+ has five parameters as content value *(C)*, dictionary threshold *(D)*, Creating Core clusters threshold *(CC)*, Expanding Clusters threshold *(EC)* and Sophisticating the clusters threshold *(S)*. Efficiency of the algorithm can change depending on these parameters on different datasets. Parameters range between 0 and 1. Figure 22 shows the representation of an individual in MOGA.

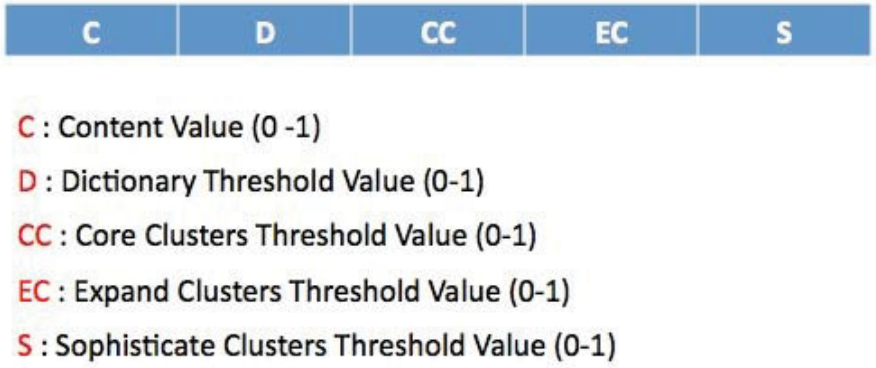


Figure 22: Individual Representation

3.2.2.2.2 Fitness Function

Four clustering objectives are defined to measure the fitness of an individual. Three of them are fixed and used in all cases desires. However, only the "Cluster Amount" objective is an optional objective, and can be chosen if the user wants. Four objectives are defined as the following:

1. *FWithin:* Represents the cohesiveness of a cluster. More cohesive clusters mean better clustering. It is estimated by calculating the average standard deviation per cluster. It measures cluster cohesiveness and shows how much deviation exists in a cluster on average. It can be defined as Eq.14:

$$FWithin = \frac{\sum_{i=1}^{m} AvgStdDev\ (Cluster_C)}{m}$$

Eq.14

2. *Fbetween:* Represents the distance between clusters. Distance between clusters means how separate clusters are from each other. More distance between clusters is better for more efficient results. *Fbetween* is estimated by using the average standard deviation of two clusters and the Euclidean distance between the word occurrence frequency vectors of the clusters. It can be defined as Eq.15:

$$FBetween = \frac{EuclideanDistanceFrom\_x\_to\_y}{\sqrt{(AvgStdDev_X)^2 + (AvgStdDev_Y)^2}}$$

Eq.15

3. *ClusteredPairAmount:* Represents the number of problem-solution pairs, which are clustered by the algorithm. A large number of clustered pairs shows that the algorithm is able to create relations between more pairs so the larger the value, the better it becomes.

4. *ClusterAmount:* Represents the number of clusters created by the algorithm. It is not possible to say if bigger ClusterAmount is better or not. This objective depends on the user's choice. A user may want to have a small number of clusters with many problem/solution pairs in them to have more general solutions, or to have larger number of clusters with a small number of problem/solution pairs to have more specific solutions. A user has to choose this objective before starting the clustering engine via the user interface shown in Figure 14.

3.2.2.2.3 Evolutionary Component

Until a certain number of epochs given, the population is evolved as in the Figure 23. Initial Population is created randomly. After calculating ranks and fitness of the individuals, two fittest individuals are chosen to be parents of new offsprings. A new offspring is generated by employing the crossover operator on the parents. Then the new offspring is mutated to have diversity in the population. Finally, the least fittest two individuals are replaced by new offsprings. After finishing the evaluation of the population, the set of non-dominated individuals are identified.

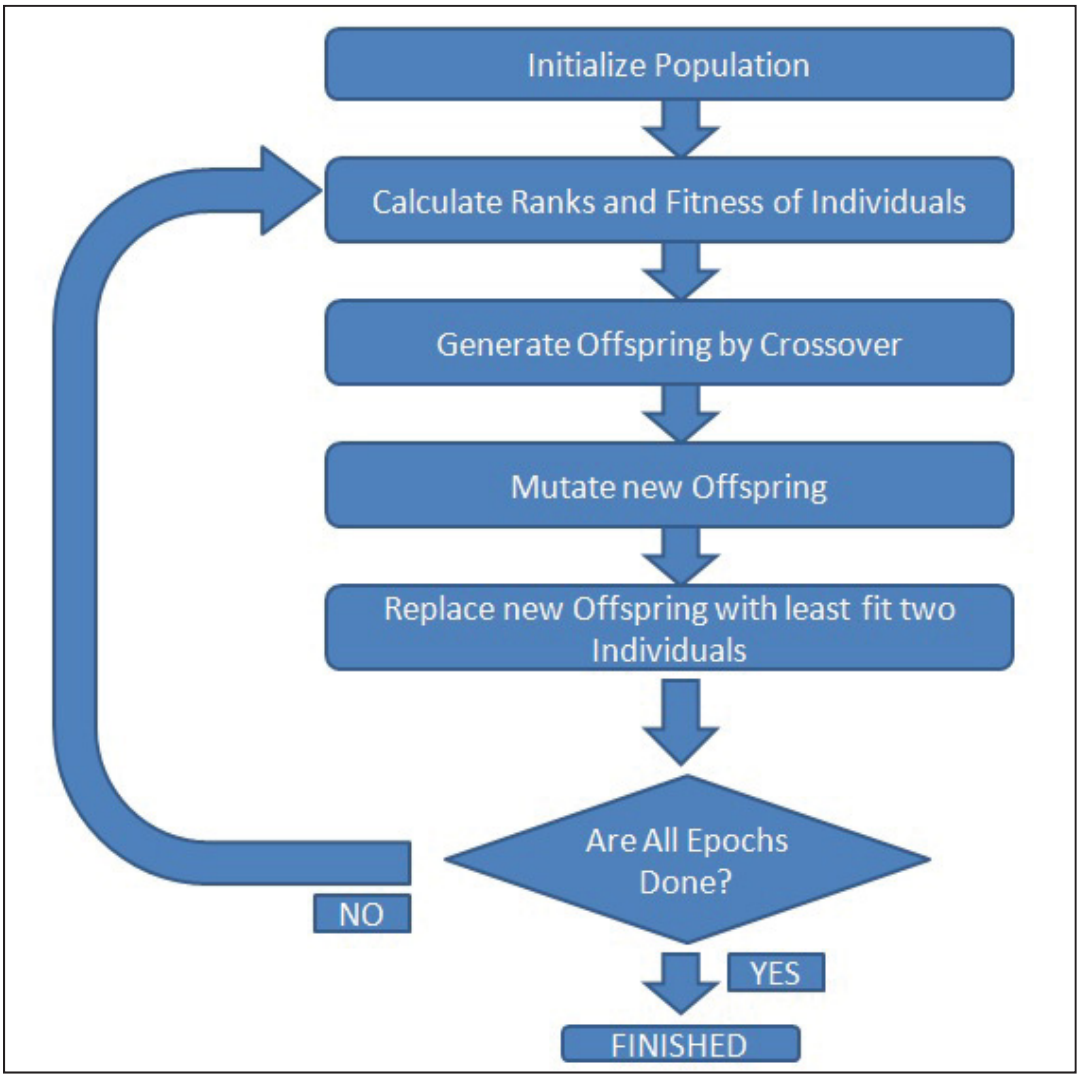


Figure 23: Evolutionary Component

3.3 Knowledgebase

The experience knowledgebase is the storage where the structured and useful information is stored. The form of knowledgebase module is shown in Figure 24.

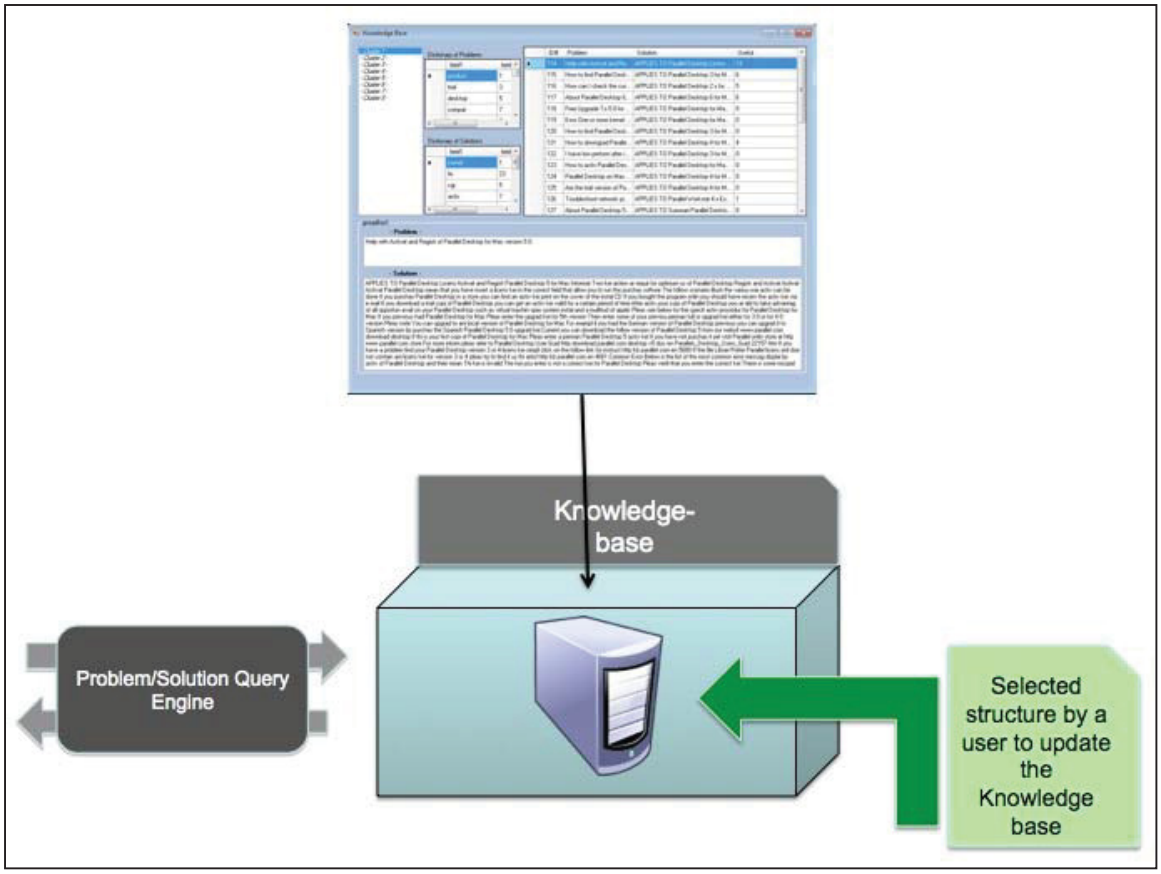


Figure 24: Knowledgebase

This module enables the user to see the current structure of the experience knowledgebase. The experience knowledgebase consists of four components as:

(i) List of Existing Clusters: Shows you the current clusters in the knowledgebase, which are created at the last system update. The following two components are used to investigate the internal structure of clusters and what they show. They change depending on the user's selection in this list. So this component enables the user to change elements of the following components.

(ii) List of Words in the Dictionaries: Enables the user to see the important words that are used to create the cluster that he/she selects from the List of Existing Clusters.

(iii) List of Problem-Solution Pairs: Shows the user the current Problem-Solution experience instances in the cluster that is selected from the List of Existing Clusters. The following component is used to investigate the internal structure of a single Problem-Solution experience instance and what it shows. This changes depending on the user's selection in this list.

(iv) Problem-Solution Pair: Shows the user the internal components of an experience instance that is chosen from the List of Problem-Solution Pairs.

3.4 Main Form and Problem/Solution Query Engine

Query engine is the right side component of the main window of the system. To send a query to the system, query pages such as pages of a web browser are used. Form window view of the main window is shown in Figure 25. A user can open a new query window by clicking on the "+" next to the query tab, and can visit between the queries by using the new query tab.

After writing a query to the “Query” box, the user can retrieve related problem-solution pairs by clicking on the “Bring Solutions” button. Top rated pairs appear on the table, but if the user would like to see all problem-solution pairs even if they have a

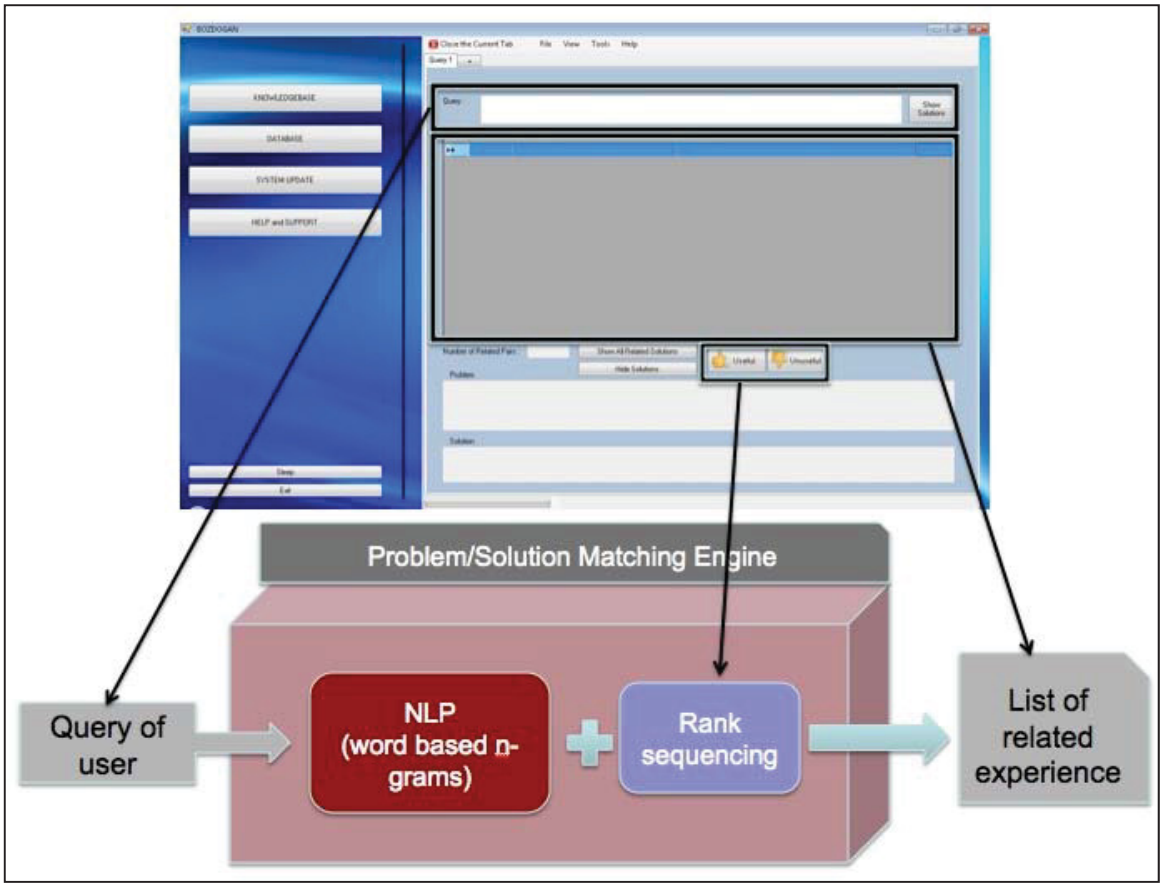


Figure 25: Main form and Problem/Solution Matching Engine

little relation with the user's query, the user can retrieve all related pairs by clicking on the "Show All Solutions" button. To see if there are any other pairs available, the user can check the "Number of Pairs" box. To define related past experiences with the user's query, word based n-grams are applied with the algorithm proposed by Keselj et al. [45]. The algorithm is as following :

Profile Dissimilarity($profile_1$, $profile_2$)
{
    sum = 0
    for all n-grams x contained in profile1 and profile2 do
        let f_1 and f_2 be frequencies of x in $profile_1$ and $profile_2$(zero if they are not included)
        add square of the normalized difference of f_1 and f_2 to sum:
        sum = sum + $(2*(f_1 - f_2)/(f_1 + f_2))^2$
    return (-sum)
}

The algorithm is measuring similarity between two profiles and in the proposed system, a profile represents experience instance. The formula used in the algorithm is defined as Eq.16:

$$\sum_{n \in \text{profile}} \left(\frac{f_1(n) - f_2(n)}{\frac{f_1(n)+f_2(n)}{2}} \right)^2 = \sum_{n \in \text{profile}} \left(\frac{2 \cdot (f_1(n) - f_2(n))}{f_1(n) + f_2(n)} \right)^2 \quad \text{Eq.16}$$

Also by using the "Useful" button, the user can rate a pair that he/she thinks it is useful for solving a problem. So next time when the user have a similar kind of problem, it will be on the top of the list depending on its usefulness degree after the calculation of n-gram similarity. On the other hand, if the user would like to decrease the usefulness degree of a pair, because he/she thinks that it is not useful for his/her query, the user can click on the unuseful button to reduce the usefulness degree of a problem/solution.

Chapter 4 - EXPERIMENTS and RESULTS

In this chapter, I am going to present the experiments and results in two subsections as Clustering Algorithm Experiments and Multi Objective Genetic Algorithm Experiments. In this thesis, coding, training, testing and all experiments are done on a desktop machine with Intel Core 2 Duo CPU E7500 @ 2.93GHz processor, 6.00GB(5.90 GB usable) RAM, Microsoft Windows 7 home premium 64-bit OS.

In the evaluation phase, the following metrics are used to compare the effectiveness of the algorithms:

Precision

It represents the rate of retrieved documents that are relevant to the actual class in that cluster.

$$Precision = \frac{|\{retrieved\ documents\}| \cap |\{relevant\ documents\}|}{|\{retrieved\ documents\}|}$$

Eq.17

Recall

It represents the rate of the retrieved documents to the relevant documents that are put into the same cluster.

$$Recall = \frac{|\{retrieved\ documents\}| \cap |\{relevant\ documents\}|}{|\{relevant\ documents\}|}$$

Eq.18

Fwithin

It is estimated by calculating average standard deviation per cluster. *Fwithin* shows how much deviation exists in clusters on average so it measures cluster cohesiveness [46]. It can be defined as in Eq.16:

$$Fwithin = \frac{\sum_{i=1}^{m} AvgStdDev(Cluster_C)}{m}$$

Eq.19

where *m* is the number of clusters. More cohesive clusters are better so smaller *Fwithin* value means better clusters.

Fbetween

It is estimated by using average standard deviations of two clusters and Euclidean distance between the word occurrence frequency vectors of the clusters [46]. It is a metric that shows how separate the clusters are from each other. *Fbetween* measure is defined as in Eq.18:

$$FBetween = \frac{EuclideanDis\tan ceFrom\_x\_to\_y}{\sqrt{(AvgStdDev_X)^2 + (AvgStdDev_Y)^2}}$$

Eq.20

As an evaluation measure, average of *Fbetween* values between all clusters are calculated. *Fbetween* measure shows the distance between the clusters so the higher *Fbetween* values represent more separate clusters.

Elapsed Time

It is the CPU processing time elapsed until the end of an algorithm. So it represents the cost of the algorithm and measured as microseconds.

Standard Deviation

It is used to show how much variation from the average exists between the features of instances. High standard deviation means there is a big range of values between features of different instances. In this experiment, standard deviation shows how related the threads in a cluster are to each other by their word frequency deviations from the mean word frequencies in the cluster. If *C* is considered as a cluster, standard deviation for *C* used in this experiment can be calculated as in Eq.19:

$$AvgStdDev(Cluster_C) = \propto xAvgStdDev(\overrightarrow{Word}_{CQ}) + \propto xAvgStdDev(\overrightarrow{Word}_{CA})$$

Eq.21

where

$$AvgStdDev(\vec{Words}_{C_Q}) = \frac{\sum_{i=1}^{m} \sqrt{\frac{\sum_{j=1}^{n} \left(Words_{Tj_{Qi}} - Mean(Words_{Tj_{Qi}})\right)^2}{n}}}{m}$$

$$AvgStdDev(\vec{Words}_{C_A}) = \frac{\sum_{i=1}^{m} \sqrt{\frac{\sum_{j=1}^{n} \left(Words_{Tj_{Ai}} - Mean(Words_{Tj_{Ai}})\right)^2}{n}}}{m}$$

where $\vec{Words}_{T_Q}$ represents question words' occurrence frequency vector of the thread *T,* $\vec{Words}_{T_A}$ represents answer words' occurrence frequency vector of the thread *T*, *n* represents the number of the threads in the cluster and *m* represents the number of the words in the problem or the solution component of the thread and

$$Mean(Word_T) = \frac{\sum_{j=1}^{n} Word_T}{n}$$

where $Word_T$ represents the word occurrence frequency and *n* is the number of threads.

Average Standard deviation represents how close the instances of the cluster are to each other on the sample space. Closer instances mean more related instances so smaller Average Standard Deviation per cluster means better clustering. This measure is considered for each cluster at the final stage individually.

4.1 Clustering Algorithm Experiments

In these experiments, the three datasets as PrincetonDS, ParallelsDS and GoDaddyDS , which are explained in details in the Methodology section before, are used as input data. Clustering algorithms namely Kmeans, EM, DBScan , CES, CES+ and the similarity measures Jaccard Distance, Cosine Similarity (explained in details in the Methodology section), are evaluated on these data sets.

To work with KMeans, EM and DBScan algorithms, data mining open source software WEKA [47] is used. Jacard Distance, Cosine Similarity, CES and CES+ algorithms are implemented in C# programming language using Microsoft VisualStudio 2010 environment. In the evaluation of the experimental work, three metrics namely precision and recall are used to compare effectiveness of all the algorithms. Additionally, for the algorithms, which are reverse engineered, namely CES, Jaccard Index and Cosine Similarity, elapsed time for giving clustering results, standard deviation of the result clusters, *Fwithin* and *Fbetween* metrics calculated to select the best results with optimum parameter values. To be able to have fair comparison of the CES algorithm with Jaccard and Cosine methods as it is described by the authors in [4], "size" is used as another metric for evaluation. However, it should be noted here that in my opinion, size should

(y axis: Fwithin and Fbetween/100)

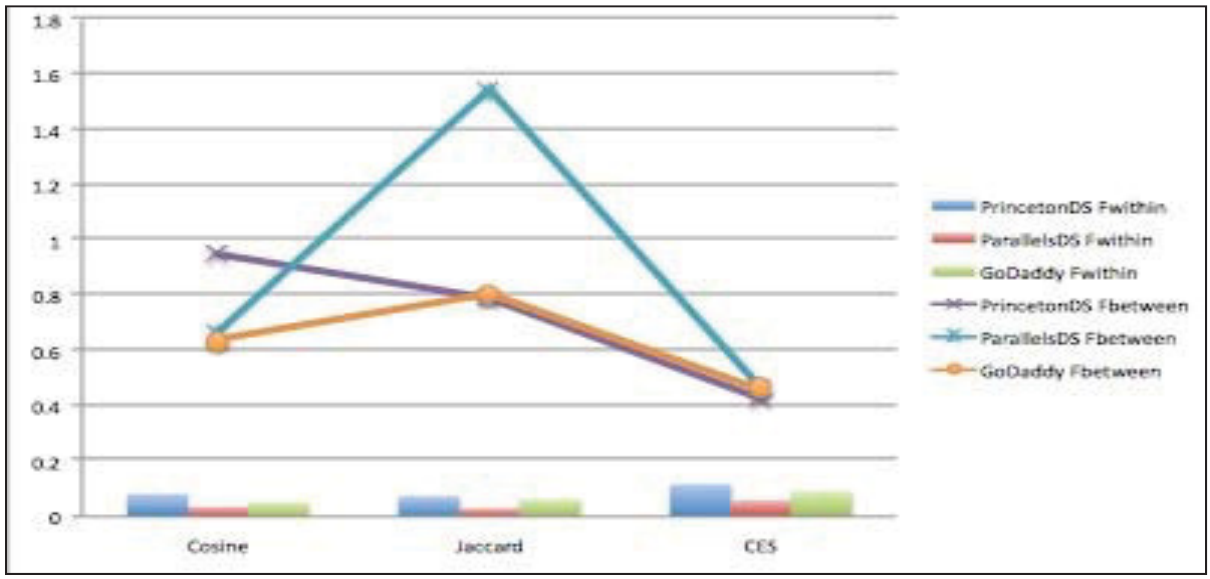


Figure 26: *Fwithin* and *Fbetween* comparisons of Cosine, Jaccard, CES on PrincetonDS, ParallelsDS, GoDaddyDS

not be considered as an evaluation metric because if the precision or recall of the cluster is low, big size is not a feature that can prove the optimality of a cluster.

Figure 26 shows the *Fwithin* and *Fbetween* results for only Cosine Similarity, Jaccard Index and CES/CES+ algorithms (in this case there is no difference between CES and CES+) on the three data sets. The reason these techniques could be analyzed by these two metrics is that because I developed the code for these techniques so the data is collected to be able to calculate these metrics, too. However, for K-Means, EM and DBScan, the open source software WEKA is employed, which does not collect data to calculate these two metrics for those algorithms. These results show that CES/CES+ performs really close to the other two on these data sets by *Fwithin* and the *Fbetween* values for clusters. This means CES+ generates cohesive and well-separated clusters, even though its vector representations are changed compared to CES.

On the other hand, the average precision and recall metrics can be calculated for all techniques on all three data sets. To this end, Figures 27, 28, and 29 show the average precision and recall for PrincetonDS, ParallelsDS, and GoDaddyDS data sets, respectively. In these figures, x-axis represents the number of clusters generated by each technique. In this case, the number of clusters that are generated by CES and CES+ techniques is very close to the real number of classes in the data sets. These techniques

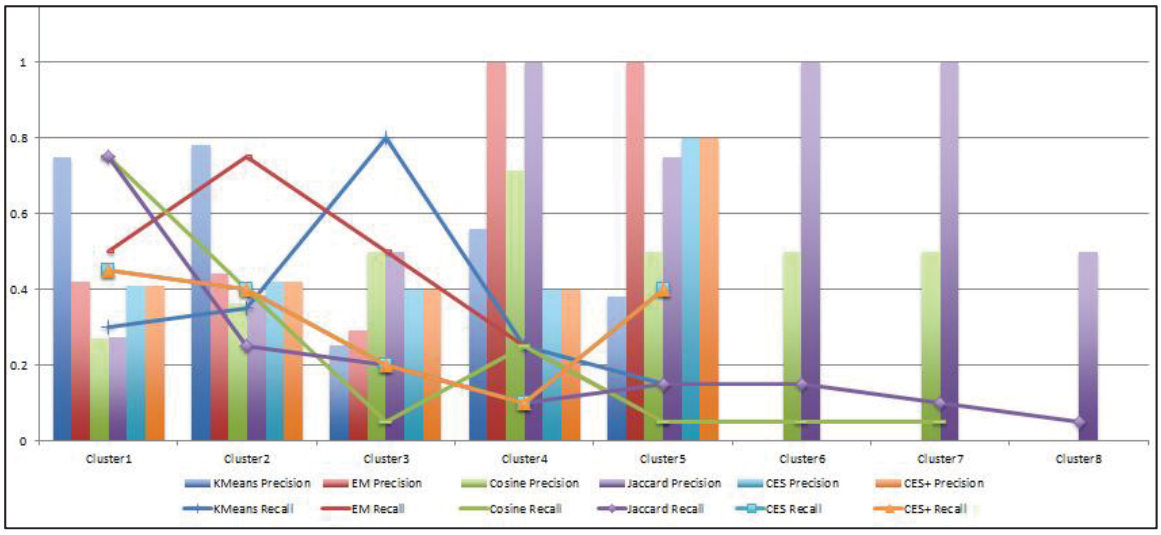


Figure 27: Precision and Recall comparisons among all employed algorithms on PrincetonDS

also give higher Precision and Recall values than the other techniques evaluated (y-axis represents values for Precision and Recall). It should be noted here that DBScan algorithm is not able to create any clusters on these datasets so it is not included in these figures.

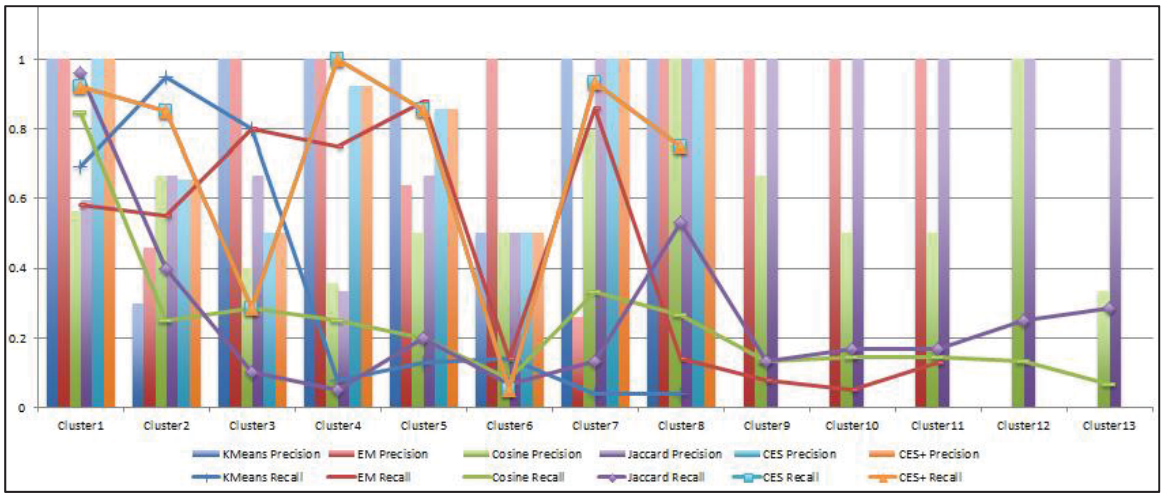


Figure 28: Precision and Recall comparisons among all employed algorithms on ParallelsDS

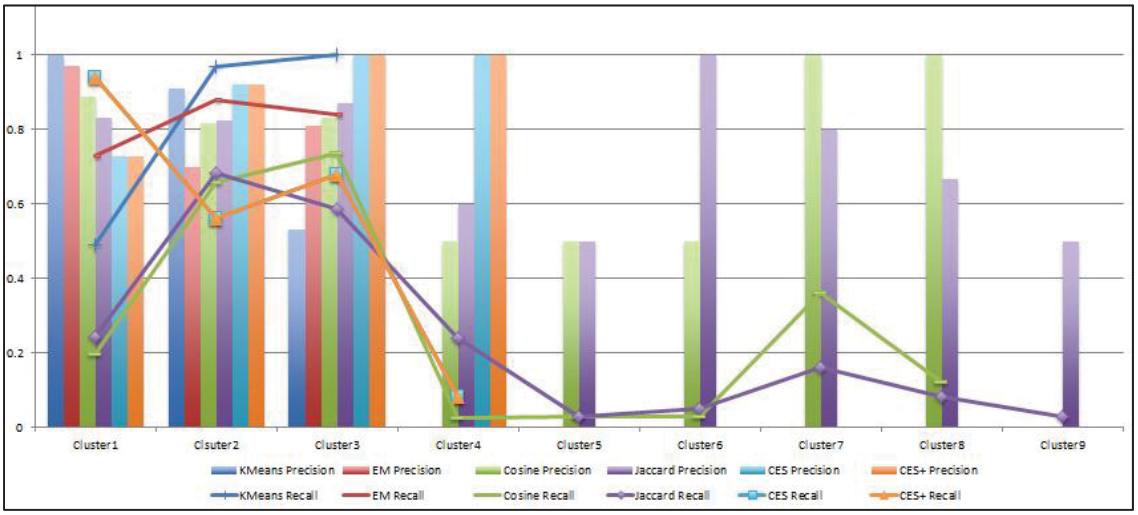


Figure 29: Precision and Recall comparisons among all employed algorithms on GoDaddyDS

Moreover, Figure 30 shows the computational cost of each algorithm on each data set based on the time (in seconds) it took for the algorithm to complete. These results show that CES+ and CES algorithms generate better clusters (low *Fwithin* and high *Fbetween*) with the highest Precision and Recall values than all the other algorithms on these data sets. However, CES+ achieves such a performance with a very low computational cost whereas CES has the highest computational cost among all the algorithms evaluated.

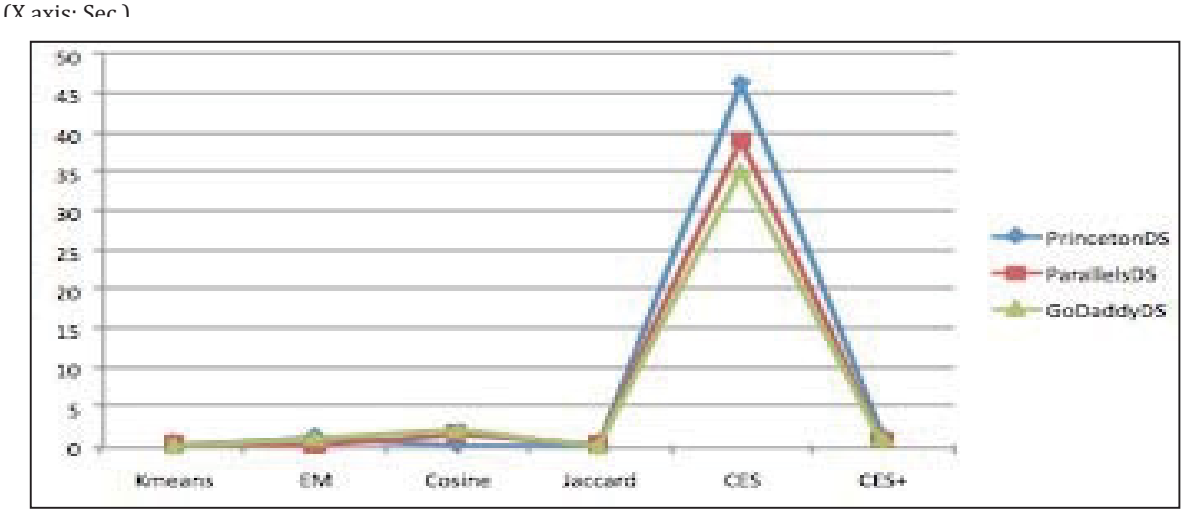


Figure 30: Time comparison among all algorithms

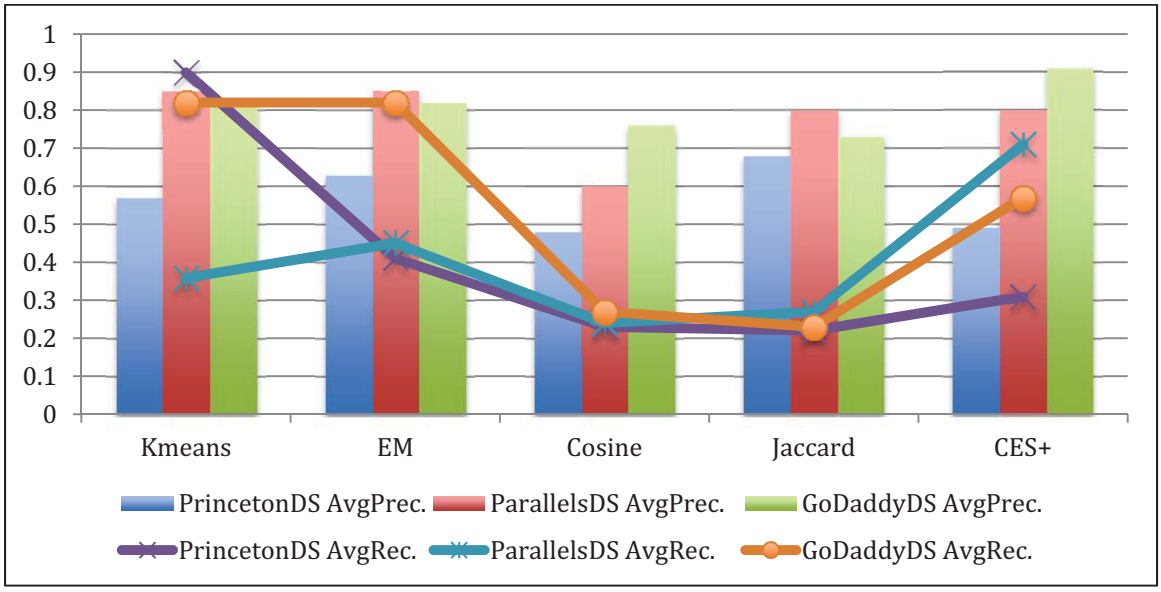


Figure 31: Average precision and recall comparisons of the algorithm results on PrincetonDS, ParallelsDS and GoDaddyDS

Additionally, Figure 31 (y axis represents values for Precision and Recall) shows average Precision and Recall comparisons among all algorithms on PrincetonDS, ParallelsDS and GoDaddyDS. Average precision and recall values are calculated by precision and recall values of all generated clusters of the algorithms. Figure 32, shows that CES and CES+ algorithms performs better than other algorithms on ParallelsDS and GoDaddyDS. On PrincetonDS, Kmeans seems to have the best performance, but if you refer to the Appendix Section 1, you will see that it generates larger number of clusters then the original number of classes and because of generating clusters with a high number of instances, it gives higher recall on average.

4.2 MOGA Experiments

For a Multi Objective Genetic Algorithm, it is important to define the number of generations to stop generating new populations when there is not a big change in the name of fitness function metrics. At the same time it is important to provide good solutions for the user to update the system.

In this subsection, the evaluations are discussed for deciding on the number of generations for the MOGA. In order to achieve this, it is experimented with 1000 generations and recorded the *Fwithin – Fbetween* values of the best pareto individuals. Figure 33, Figure 36, Figure 39 present the actual *Fwithin* values of individuals for PrincetonDS, ParallelsDS, GoDaddyDS, respectively. Figure 34, Figure37, Figure 40 present the actual *Fbetween* values of individuals for PrincetonDS, ParallelsDS, GoDaddyDS, respectively. Figure 35, Figure 37 and Figure 41 present normalized vlaues of *Fwithin* and *Fbetween* to make it possible to compare them on the same figure. For more detailed results, you can refer to the Appendix Section 2.

4.2.1 PrincetonDS

X axis represents the number of generations and Y axis represents the value of *FBetween* in Figure 32.

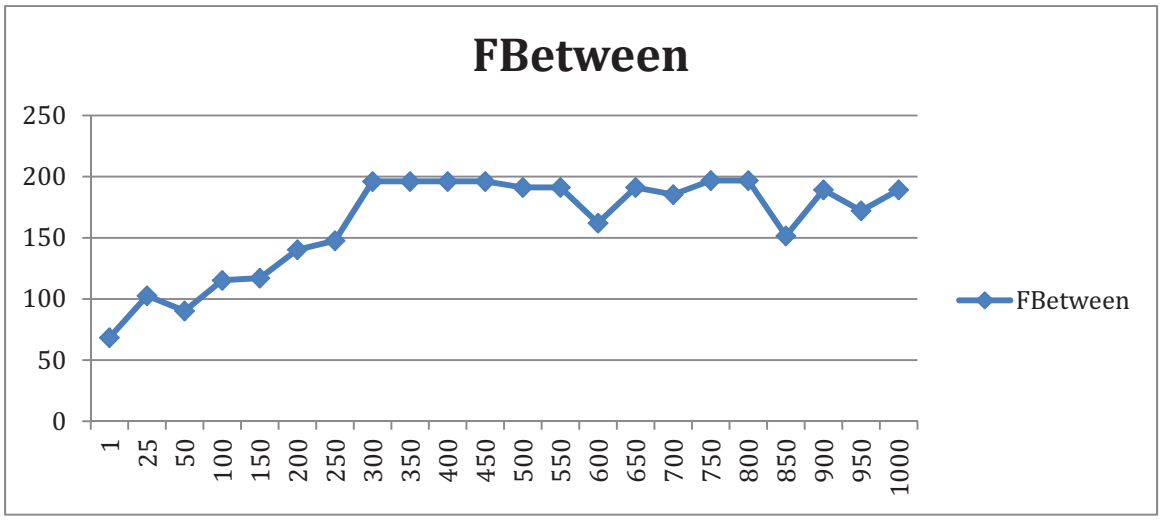


Figure 32: *Fbetween* comparison among generations on PrincetonDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* in Figure 33.

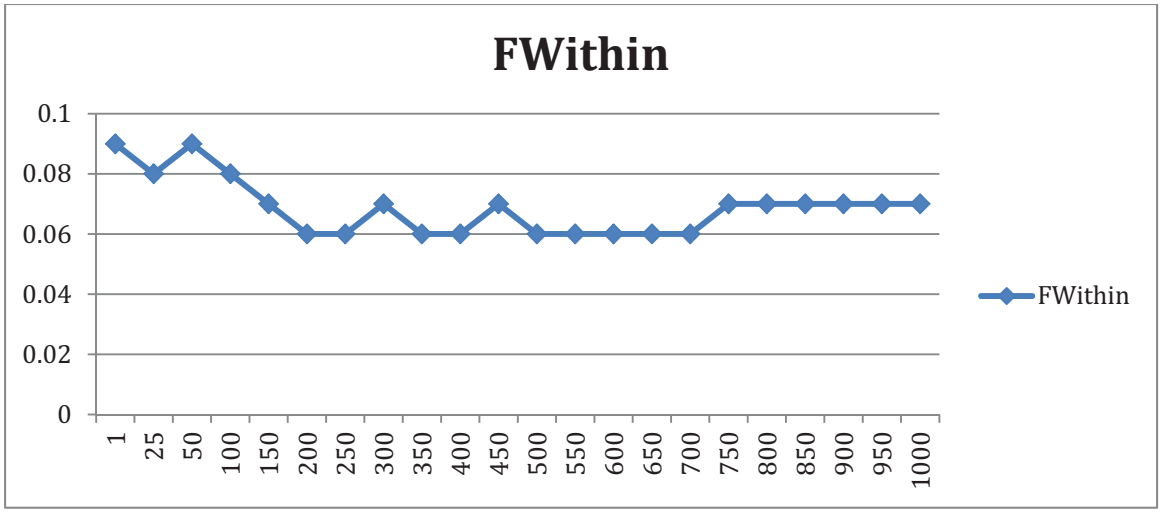


Figure 33: *Fwithin* comparisons among generations on PrincetonDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* and (the value of *Fbetween*)/1000 in Figure 34.

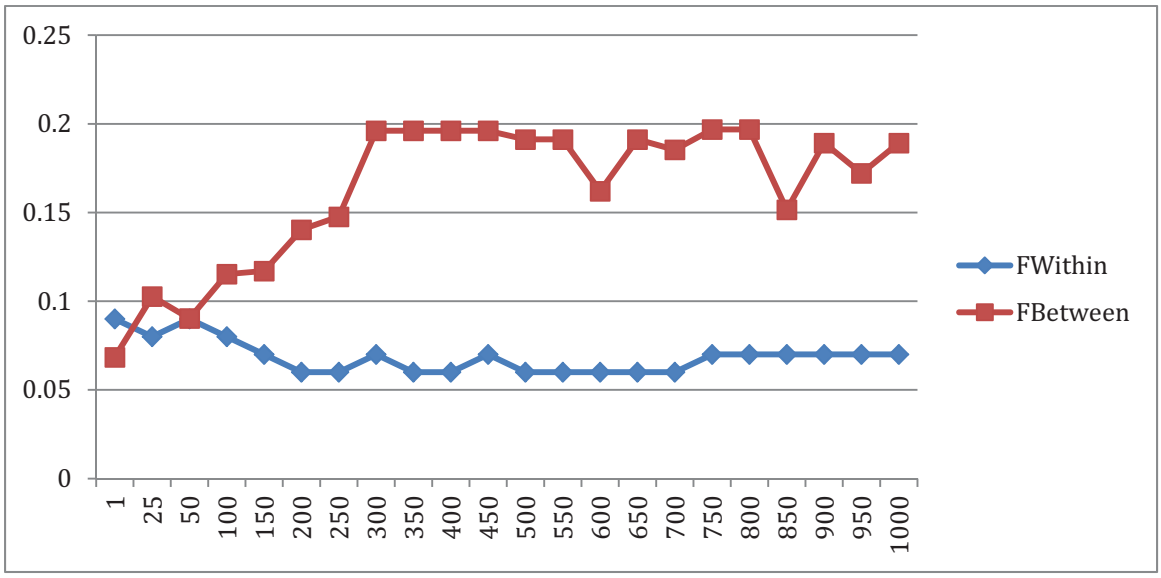


Figure 34: *Fwithin* and *Fbetween*/1000 comparison on generations

4.2.2 ParallelsDS

X axis represents the number of generations, Y axis represents the value of *FBetween* in Figure 35.

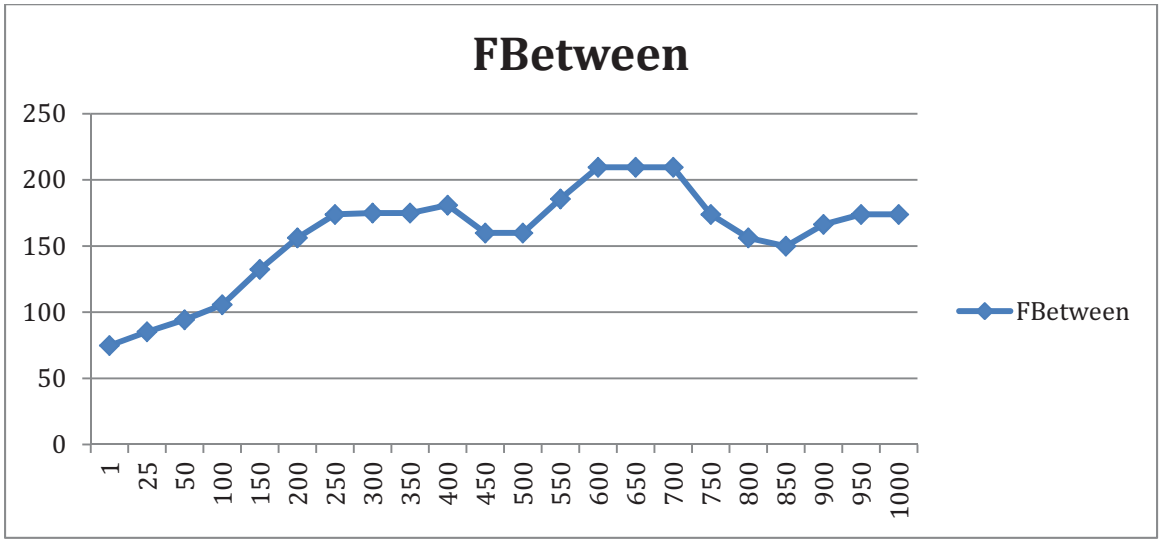


Figure 35: *Fbetween* comparison among generations on ParallelsDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* in Figure 36.

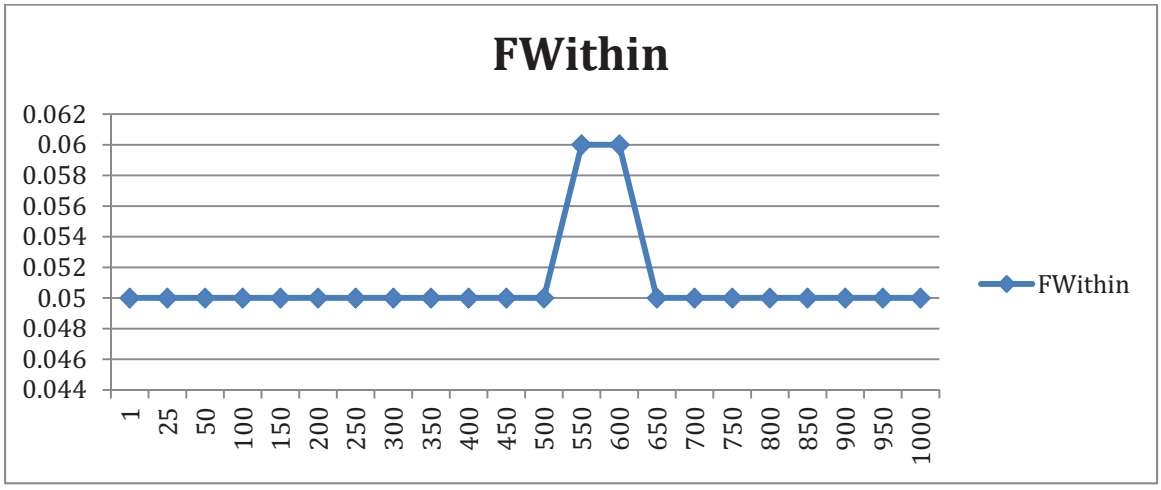


Figure 36: *Fwithin* comparisons among generations on ParallelsDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* and (the value of *Fbetween*)/1000 in Figure 37.

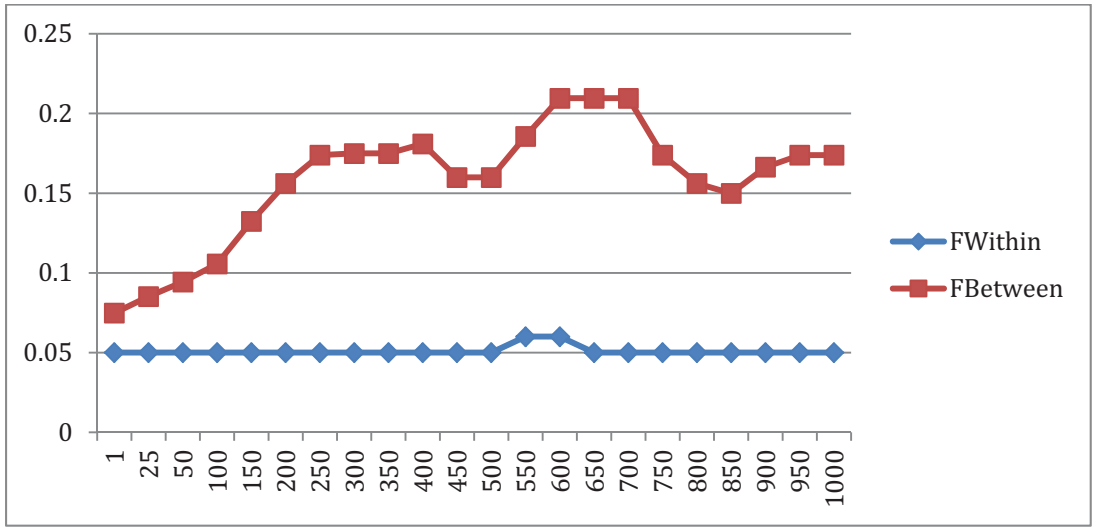


Figure 37: *Fwithin* and *Fbetween*/1000 comparisons among generations on ParallelsDS

4.2.3 GoDaddyDS

X axis represents the number of generations, Y axis represents the value of *FBetween* in Figure 38.

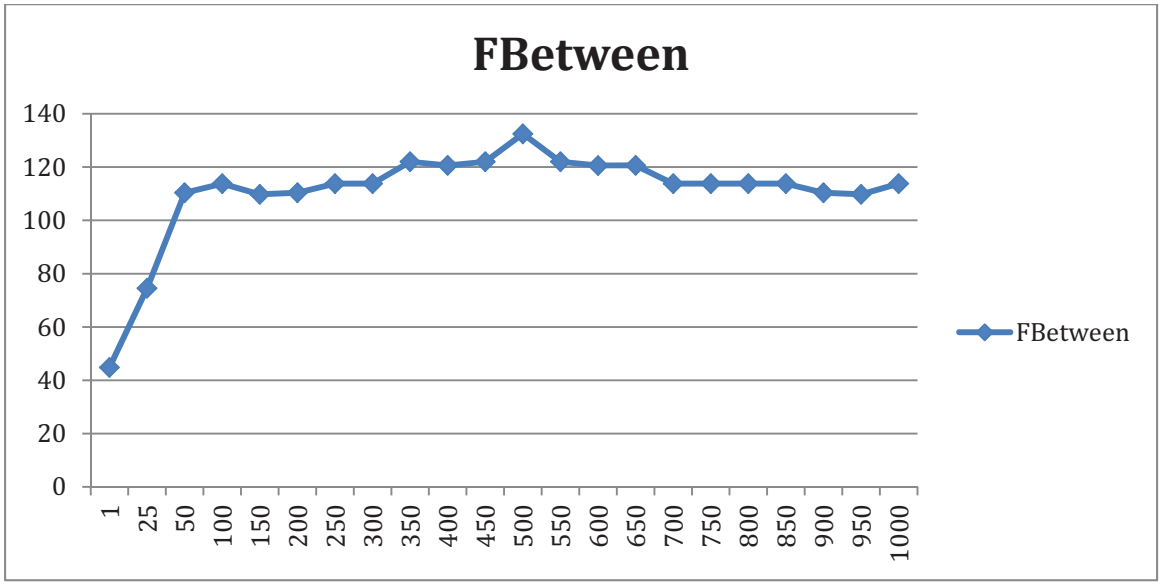


Figure 38: *Fbetween* comparisons among generations on GoDaddyDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* in Figure 39.

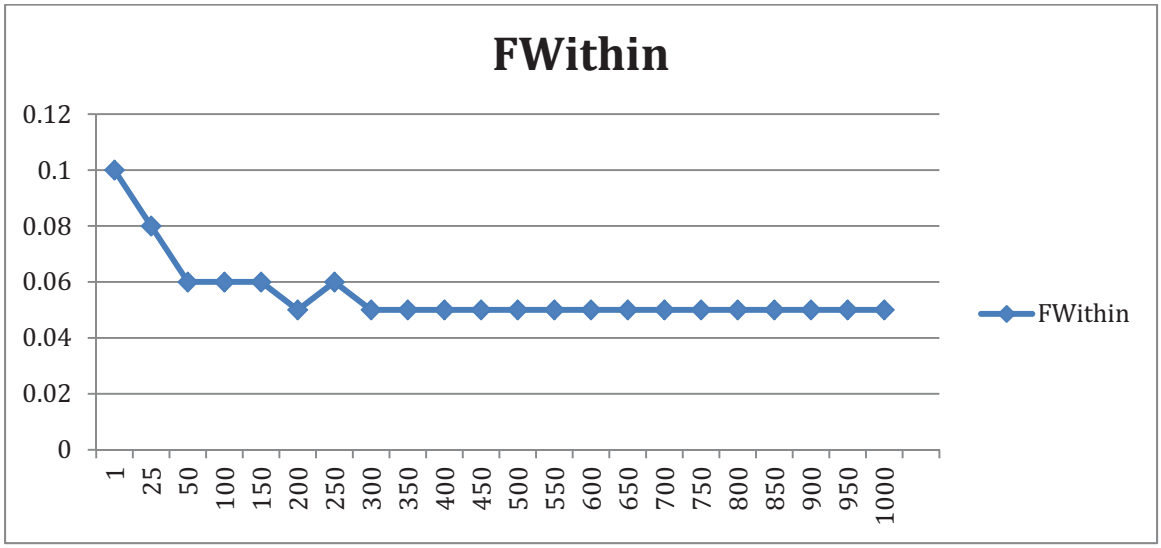


Figure 39: *Fwithin* comparisons among generations on GoDaddyDS

X axis represents the number of generations and Y axis represents the value of *Fwithin* and (the value of *Fbetween*)/1000 in Figure 40.

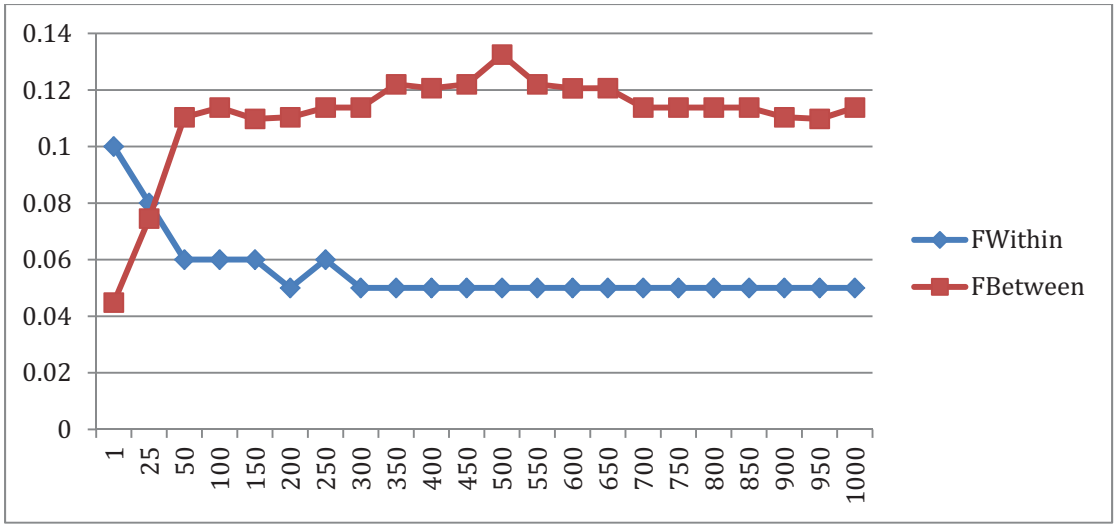


Figure 40: *Fwithin* and *Fbetween*/1000 comparison among generations on GoDaddyDS

As it can be seen from the figures, on the average, 500 generation is the optimum for MOGA to reach high *Fbetween* and low *Fwithin* values for the employed datasets. So in all experiments with MOGA, the generation number is defined as 500.

4.3 CES+MOGA Experiments

Cross validation is a model evaluation method to overcome the problem of realizing the biasing of clustering algorithms. Sometimes evaluating an algorithm with a dataset can not give an indication of how well it will perform when a first time seen data set is used. If the algorithm is not able to work with a new dataset, which it has not seen before, that means it may not be a practical method to use in real life applications where the likelihood of seeing different data sets is high. One way to overcome the problem of realizing how practical an algorithm is to not use the entire dataset when training the learning algorithm. Some of the data is removed from the dataset and it is used to test the algorithm after training.

In this thesis, three data sets are employed to evaluate the performance of the proposed system on different data sets. Moreover, because the three data sets are relatively small, N-fold cross validation technique is used to avoid any bias that may exist in these data sets. In N-fold cross validation, the data set is divided into *N* subsets. The training and testing is applied *N* times while each time one of the *N* subsets is used as the test set and N-1 subsets are used for training. Then the average error across all N trials is computed. Figure 41 shows the graphical representation of cross validation method where boxes represent the N subsets and specifically, black boxes represent the test data.

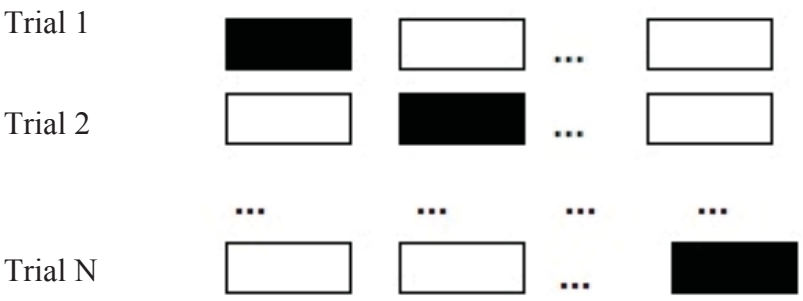


Figure 41: Graphical representation of cross validation algorithm

In all the experiments performed to test CES+MOGA, 10 fold cross validation is used. The dataset is divided into 10 subsets and then the system is trained and tested five

times on each sub dataset, each of these 5 is a 10-fold cross validation. Firstly, CES+MOGA is trained on training data, then the best individual solution of the GA is chosen for identifying the parameters for CES+ and finally CES+ is run with these parameters on the test data. In total, 50 experiments are run on each data set. Figure 42 shows the average results of these experiments on the three data sets with 10-fold cross validation method. For more detailed results, you can refer to Appendix Section 3.

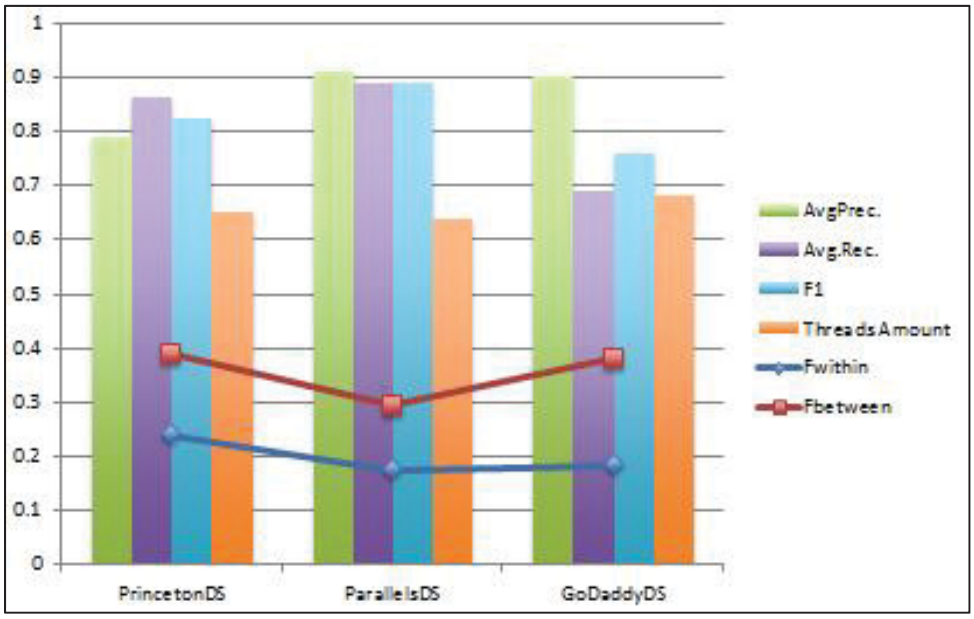


Figure 42: Results of 10-fold cross validation on three different datasets

Results show that CES+MOGA performs good on different datasets with its high average precision, average recall, average *Fwithin*, average *Fbetween* and average F1 values. The results show that the proposed model (CES+MOGA) can be used effectively in practice on real life data sets.

4.4 A Sample Case

In this section, a sample case is presented to show how the proposed system can be used in real life cases. A query is sent to the proposed system by the UI developed. The system shows the most similar 5 experiences but a user can also see other experiences depending on his/her preferences.

As a real life sample case scenario, consider a company that newly started to use Parallels products. Also the company uses the proposed system for the support of its IT management team. Consider that IT management team collected experience data about Parallels products after the management team of the company decided to use Parallels products a while ago. The experience data is collected from publicly available web source "kb.parallels.com" by using web crawler engine and web page parser components of the proposed system.

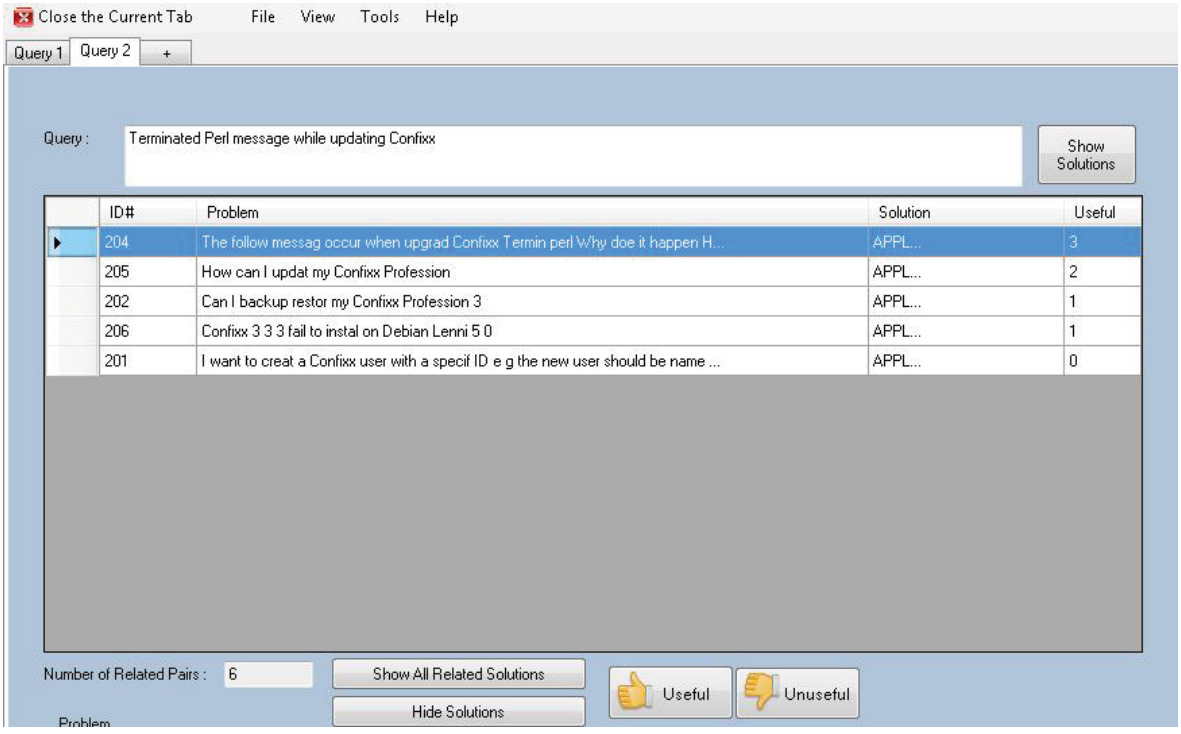


Figure 43: Sent query and retrieved solutions of a sample Case

So database of the proposed system contains experience data about Parallels products. Thus, the dynamic knowledgebase of the proposed system contains also knowledge about

the Parallels products. One day, a member of the IT management team receives an email, which tells that an error message as “Terminated perl” is received while an employee is trying to update the Confixx software in his/her laptop. In this case, the IT management team member uses the proposed system to get fastest support to solve that problem and he sends a query as “Terminated perl message while updating Confixx” to the system as shown in the Figure 43. Then he receives 5 most similar experiences for that problem in a second and replies back to the employee in shortest time with the solution by using most helpful experiences for that case. For the advantage of next searches, the proposed system also allows users to rate the retrieved experiences for sent queries by using “Useful” button to increase the usefulness rate and “Unuseful” button to decrease the usefulness rate. By this way, the system considers the usefulness of the experiences in the system for the next query and retrieval processes. The usefulness rate of experiences is used to define, which experience should be shown on top, when the similar experiences are brought to the user as solutions as explained in section 3.4.

The retrieved solutions of the proposed scenario are provided below. To see them by using the UI of proposed system, a user should focus on the retrieved experiences one by one.

Sent Query: Terminated perl message while updating Confixx

Retrieved experiences:

Problem

The following messages occur when upgrading Confixx: Terminated perl. Why does it happen? How can I solve the problem?

Solution

If there are many users in your Confixx or there are not enough hardware resources, some upgrading tasks can be closed by the system due to the lack of resources. That is why these messages show up. To resolve this situation you have to wait until upgrade process ends and then launch the terminated scripts manually. Replace the $BASE in the message with your Confixx installation directory (typically /root/confixx) to specify the location

of a script. Note, that it can take quite a long time to execute these scripts. If you want to avoid this problem in future upgrades you can do the following: Make sure that the confixx_counterscript.pl and confixx_updatescript.pl files will not be launched while upgrade is in progress. Comment out the Confixx related records in your crontab. Stop FTP and Apache servers before starting upgrade. Stop another resource wasting services (except MySQL), if any.

Problem

How can I update my Confixx Professional?

Solution

You should download a Confixx Professional update and an upgrade instruction (release_notes) from our website:
http://www.swsoft.com/en/download/confixx/confixx31

Problem

Can I backup/restore my Confixx Professional 3?

Solution

Confixx Professional 3 has Backup and Restore utilities for that: backup.pl restore.pl These utilities can be found in the Confixx installation directory (it is /root/confixx by default).

The Confixx Backup utility saves the content and settings of the entire Confixx server in a single file. This file can then be restored with the help of the Confixx Restore utility (restore.pl).

The Confixx Restore utility restores the content and settings of the entire Confixx server from a special backup file. This file is created with the help of Confixx Backup utility (backup.pl). You can get more information on these utilities with help of "-h" option. #./backup.pl –h #./restore.pl –h An example of Confixx Backup utility: #./backup.pl --dump dump

After execution of this command you will have the backup of your Confixx Professional: dump.tgz

An example of Confixx Restore utility: #./restore.pl --dump dump --map my_map.map -mapping –clean

Edit the 'my_map.map' file as desired.

#./restore.pl --dump dump --map my_map.map --restore --clean

Problem:

Confixx 3.3.3 fails to install on Debian Lenny (5.0)

Solution:

Before the Confixx installation on Debian Lenny (5.0), you need to check apache configuration first. Open apache config file /etc/apache2/apache2.conf

If it contains such lines:

PidFile ${APACHE_PID_FILE}

User ${APACHE_RUN_USER}

Group ${APACHE_RUN_GROUP}

then replace them with

PidFile /var/run/apache2.pid

User www-data

Group www-data

Save the changes and restart apache.

Now you can install Confixx.

Problem:

I want to create a Confixx user with a specific ID (e.g the new user should be named web33). How can I do it?

Solution:

When you are filling the personal data of a Confixx user you need to enter "add-user" in the third "Definable field" (left input) and the required ID in the corresponding "Input" field (right input). If this ID has not been taken by another Confixx user yet it will be assigned to your Confixx user.

Chapter 5 – CONCLUSION and FUTURE WORK

The objective of this research is to minimize the overall expense of IT fault management by providing the right mix of automation and manual activity by proposing a system support decision making process for IT problem solving tasks. The identification of the root cause and planning the resolution for the problems/failures in an IT system are the biggest concerns that have high costs. The proposed approach utilizes IR and data mining techniques and uses CES+MOGA to automatically extract information from publicly available resources such as helpdesk web sites FAQs. Thus, the main motivation behind this research is to automatically generate a knowledgebase for supporting IT management teams. The use of decision support tools, trouble shooting systems, incident management tools etc. for IT management support has already been demonstrated by various authors reviewed in chapter 2. However the main contribution of this thesis comes from proposing a system that can support small to medium sized companies or new organizations where having an experienced IT personal is an issue because of higher cost. So, a system that can automatically provide much needed experience to solve problems may be valuable in such environments.

The first part of this thesis research consists of investigating whether an IR approach with a clustering algorithm may perform efficiently on generating experience knowledgebases from publicly available experience data on the Internet. Chapter 3 addresses this by benchmarking the performance of two different well known distance measures namely Cosine Similarity Measure, Jaccard Index and five different well known clustering algorithms namely K-Means, EM, DBScan, CES, CES+. Three different datasets, which are generated from publicly available well known web pages, namely PrincetonDS, GoDaddyDS, ParallelsDS are employed for the evaluations. Based on the *Fwithin*, *Fbetween*, Precision, Recall and computational cost (time) performance metrics, CES+ performed better than other techniques on these datasets. The results show that CES+ is an effective algorithm to be used in the proposed system.

The second phase of this work is to investigate whether it is possible to provide automated parameter optimization of CES+ for real time usage for the proposed system by using MOGA in order to automatically identify the best parameters for CES+ given a

publicly available data set. In chapter 3, CES+MOGA approach is evaluated on the aforementioned datasets. In this case, the datasets are used to compare manually configured CES+ and automatically configured CES+MOGA results. Additionally, 10 fold cross validation method is applied to evaluate the performance of CES+MOGA on unseen datasets. Evaluations are made based on the *Fwithin*, *Fbetween*, Precision and Recall performance metrics. The results show that MOGA support is not only automating the parameter optimization of CES+, it also provides more efficient results in terms of defined performance metrics.

In summary, the proposed system showed promising results in terms of generating an experience knowledgebase from publicly available data sets for the IT management support. Automation from retrieving publicly available online data to generating a dynamic knowledgebase, is the approach that is employed in this thesis. It is important to see that the balance of automation and manual activity is a vital point to develop systems for real life usage for providing effective IT management support. Experimental results indicate that the usage of IR and data mining techniques for automation can be a good model for IT management support for small to medium sized organizations and companies. The following presents some directions for the future work:

- With regards to the web crawler that is used, different parameter configurations can be made to gather more experience data from different web sites. Moreover, different web crawlers can be used to collect publicly available data depending on the preference of the user.
- With regards to the web page parser, improvements such as better stemming, better stop word removing etc. can be introduced for obtaining better results in generating structured database files from crawled web page folders.
- Additionally, more datasets and bigger data sets from different sources can provide better test environments for the proposed system.
- With regards to the CES+ algorithm, improvements in the steps of the algorithm such as comparing different methodologies in different steps may lead for better results. Moreover, using different n-grams may result

with better output of CES+ and better support of problem-solution matching engine.

- Finally, different user interfaces can be developed for human computer interaction tests and usability of the system can be tested in a real time environment by the usage of IT management team members in an organization.

Bibliography

[1] Kirmani E., Hood C. S., Diagnosing network states through intelligent probing, IEEE/IFIP Symposium on Network Operations and Management Symposium, pp. 147 – 160, 2004.

[2] Wong A.K.Y, An C. C., Paramesh N., Rav P., Ontology mapping for network management systems, IEEE/IFIP Symposium on Network Operations and Management Symposium, pp. 885 – 886, 2004.

[3] Bergmann R., Experience management: foundations, development methodology, and Internet based applications, LNCS, Springer, 2002.

[4] Iwai, K.; Iida, K.; Akiyoshi, M.; Komoda, N.; "A help desk support system with filtering and reusing e-mails," 8th IEEE International Conference on Industrial Informatics, pp. 321-325, 2010.

[5] Bartolini, C., Stefanelli, C., "Business-driven IT management," IFIP/IEEE International Symposium on Integrated Network Management, pp.964-969,2011.

[6] Bartolini, C., Stefanelli, C., Tortonesi, M.; "On decision making in business-driven IT management," IFIP/IEEE International Symposium on Integrated Network Management, pp.1082-1088, 2011.

[7] Li H., Zhan Z., “Bussiness-Driven Automatic IT Change Management Based on Machine Learn- ing”, the 7th IFIP/IEEE International Workshop on Business-Driven IT Management (BDIM) in con- junction with IFIP/IEEE NOMS,2012 - Maui,Hawaii,USA

[8] Marcu, P., Schaaf, T., "An information model for inter-organizational fault management," Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on , vol., no., pp.1043-1049, 23-27 May 2011

[9] Bartolini, C., Stefanelli, C., Tortonesi M., "Business-impact analysis and simulation of critical incidents in IT service management", in Proceedings of the II'h IFIP/IEEE International Symposium on Integrated Network Management (1M 2009), pp.9-16, 1-5 June 2009, New York, NY, USA.

[10] Bartolini C., Stefanelli C., M. Tortonesi, "SYMIAN: Analysis and Performance Improvement of the IT Incident Management Process", IEEE Transactions on Network and Service Management, Vol.7, No.3, pp.132-144, September 2010.

[11] Bartolini C., Stefanelli C., Targa D., Tortonesi M., ”A Cloud-based Solution for the Performance Improvement of IT Support Organizations”, the 7th IFIP/IEEE Mini Conference

[12] George A., Makanju A., Zincir-Heywood A. N., Milios E., “Information Retrieval in Network Administration”, IEEE Conference on Computer Networks and Services Research, pp. 561-568, 2008.

[13] dos Santos, R.L.; Wickboldt, J.A.; Lunardi, R.C.; Dalmazo, B.L.; Granville, L.Z.; Gaspary, L.P.; Bartolini, C.; Hickey, M.; , "A solution for identifying the root cause 82 of problems in IT change management," Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on , vol., no., pp.586-593, 23-27 May 2011.

[14] Xie X., Zhang W., Xu L., , "A lightweight description model to support experience management," Intelligent Agent Technology, IEEE/WIC/ACM International Conference on , vol., no., pp. 694- 697, 19-22 Sept. 2005 doi: 10.1109/IAT.2005.144
[15] Wong A.K.Y, An C. C., Paramesh N., Rav P., Ontology mapping for network management systems, IEEE/IFIP Symposium on Network Operations and Management Symposium, pp. 885 – 886, 2004.
[16] Melchiors C., Tarouco L. M. R., Troubleshooting network faults using past experience, IEEE/IFIP Symposium on Network Operations and Management Symposium, pp. 549 – 562, 2000.
[17] Burgess J., Guillermo R., Raising network fault management intelligence, IEEE Symposium on Network Operations and Management, 2000.
[18] WebSPHINX : A Personal, Customizable Web Crawler Available at http://www.cs.cmu.edu/~rcm/websphinx/ December,2011
[19] Princeton University Office of Information Technology Knowledgebase. Available: http://helpdesk.princeton.edu/
[20] Parallels. Virtualization and automation Software Parallels Knowledgebase. Available at: http://kb.parallels.com/
[21] Go Daddy, Go Daddy Help Center. Available at: http://help.godaddy.com/
[22] Princeton University Office of Information Technology Knowledgebase, FAQ page. Available at: http://helpdesk.princeton.edu/kb/display.plx?ID=9589
[23] Porter Stemming Algorithm. Available at: http://tartarus.org/~martin/ PorterStemmer/
[24] Parallels. About page. Available at: http://www.parallels.com/ca/about/
[25] Parallels. Virtualization and automation Software Parallels Knowledgebase., FAQ page. Available at: http://kb.parallels.com/en/111206
[26] Parallels. Virtualization and automation Software Parallels Knowledgebase, FAQ page. Available at: http://help.godaddy.com/article/328?locale=en
[27] Borman S., The Expectation Maximization Algorithm – A short tutorial, Tutorial Notes, July 2004. Available at: http://www.seanborman.com/ publications/
[28] MacQueen, J.B. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.
[29] Bilmes J., A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, 1988. Available at: lasa.epfl.ch/teaching/lectures/ML_Phd/Notes/GP-GMM.pdf
[30] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, " A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", The 83 Second International Conference on Knowledge Discovery and Data Mining (KDD- 96), Portland, Oregon, USA, 1996
[31] Moreira A., Santos M.Y., Carneiro S.,"Density-based clustering algorithms – DBSCAN and SNN", University of Minho, Portugal, 2005
[32] Tan P., Steinbach M., Kumar V.,"Cluster Analysis: Basic Concepts and Algorithms" in Introduction to Data Mining, Addison-Wesley, 2005
[33] Anderberg, M.R., Cluster Analysis for Applications. Academic Press, 1973

[34] Alpaydin E., Introduction to Machine Learning. MIT Press, 2004.
[35] Alsabti K., Ranka S., Singh V., "An efficient k-means clustering algorithm", Electrical Engineering and Computer Science. Paper 43, 1997
[36] Rahman M., Hassan M.R., Buyya R., "Jaccard Index based availability prediction in enterprise grids", presented at Int. Conf. on Membrane Computing, pp.2707-2716, 2010
[37] Mitchell M., An Introduction to Genetic Algorithms, MIT Press Cambridge, 1996
[38] Coello Coello, C.A.; , "Evolutionary multi-objective optimization: a historical view of the field," Computational Intelligence Magazine, IEEE , vol.1, no.1, pp. 28- 36, Feb. 2006
[39] Kalyanmoy D., Multi-Objective Optimization using Evolutionary Algorithms, Berkley, CA, 2002 pp. 1-5
[40] Fonseca C.M., Fleming P.J., "Genetic algorithms for multiobjective Optimization: Formulation, discussion and generalization," In Stephanie Forrest, editor, Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 416–423, San Mateo, California, University of Illinois at Urbana-Champaign, Morgan Kauffman Publishers, 1993.
[41] Srinivas N., Deb K., "Multiobjective optimization using nondominated sorting in genetic algorithms," Evolutionary Computation, vol. 2, no. 3, pp. 221–248, Fall 1994.
[42] Horn J., Nafpliotis N., Goldberg D.E., "A niched pareto genetic algorithm for multiobjective optimization," In Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence, vol. 1, pp. 82–87, Piscataway, New Jersey, IEEE Service Center, June 1994.
[43] Kim Y., Street W.N., Menczer F, "Feature selection in unsupervised learning via evolutionary search", in KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 365–369, New York, NY, USA, 2000. ACM.
[44] Bacquet C., Zincir-Heywood A.N., Heywood M.I., "An investigation of multiobjective genetic algorithms for encrypted traffic identification," International Workshop on Computational Intelligence in Security for Information Systems, vol. 63,2009. 84
[45]] Keselj V., Peng F., Cercone N., Thomas C., "N-Gram Based Author Profiles For Authorship Attribution", Pacific Association for Computational Linguistics, pp. 255- 264, 2003.
[46] Bacquet C., Zincir-Heywood A.N., Heywood M. I., "Genetic Optimization and Hierarchical Clustering applied to Encrypted Traffic Identification", IEEE Symposium on Computational Intelligence on Cyber Security, pp. 194-201, 2011.
[47] WEKA tool. Available at: http://www.cs.waikato.ac.nz/ml/weka/.

APPENDIX SECTION 1 – CLUSTERING ALGORITHM and DISTANCE MEASURE RESULTS

Subsections 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7 contain information and explanation about the experiments. For the result tables, please refer to the Subsection 1.8.

1.1 KMeans Results

For the parameters of KMeans algorithm, default values are defined as 500 for maximum iteration number and 10 for seed number. On the tables, “InC. Classi.” represents the incorrectly classified instance number depending on the “Considered As” section. Values of the parameters are defined on the right side where “N:” represents the numer of clusters, “A:” represents the distance algorithm, “I:” represents the maximum iteration number, “S:” represents the seed number. “NoClass” phrase under "ConsideredAs" section means that Weka did not attend any class name for that cluster, because the class label is attended before. But having clusters with the same “ConsideredAs” class is acceptable, because of that, the considered class number, which has most population in that cluster, is added after the “-” symbol for the “ConsideredAs” section of the table. Three datasets, which are made by using “only question”, “only answer” and “both question and answer” parts of the question-answer pairs are used in KMeans tests for each of PrincetonDS, ParallelsDS and GoDaddyDS data sets. Searching for the optimum value for parameters is done as in N, A and S order. Firstly the optimum value for N is searched in all datasets, then depending on the results, the parts of the problem-solution pairs are defined to be used in clustering. Then the experiments go by searching for an optimum value for the next parameter following the aforementioned order. Used parameter values and results of the KMeans algorithm on the datasets are given below.

1.1.1 PrincetonDS

N is set to 2, 3, 4, 5, 6, 7 respectively while the distance algorithm is set to Eucledian Distance (E) and then to Manhattan Distance (M) and other parameters are set to default values. Then with the optimum N and distance measure, S is set to 5,6,7,8,9,10,11,12,13,14,15 respectively.

Best results are received while N is set to 5 and A is set to M while Only Question dataset is used as shown in Table 1. So these two parameter values are used to find optimum seed number S. Best results are received with the highest and the most balanced precision, recall values and one of the lowest "InC. Classi." value while S is set to 13 on the "Only Question" dataset as shown in Table 2. Results show that best results of KMeans on PrincetonDS are received by setting parameters as N:5 A:M I:500 S:13 and using "Only Question" data.

1.1.2 ParallelsDS

N is set to 5, 6, 7, 8, 9, 10, 11 respectively while the distance algorithm is set to Eucledian Distance(E) and then to Manhattan Distance(M) and other parameters are set to default values. Then with the N and distance measure, S is set to5,6,7,8,9,10,11,12,13,14,15 respectively.

Best results are received while N is set to 8 and A is set to E while "Only Question" dataset is used as shown in Table 3. So these two parameter values are used to find optimum seed number S. Best results are received with the highest and the most balanced precision, recall values and one of the lowest "InC. Classi." Value while S is set to 10 on the "Only Question" dataset as shown in Table 4. Results show that best results of KMeans on ParallelsDS are received by setting parameters as as N:8 A:E I:500 S:10 an using "Only Question" data.

1.1.3 GoDaddyDS

N is set to 2, 3, 4, 5 respectively while the distance algorithm is set to Eucledian Distance(E) and then to Manhattan Distance(M) and other parameters are set to default values. Then with the N and distance measure, S is set to5,6,7,8,9,10,11,12,13,14,15 respectively.

Best results are received while N is set to 3 and A is set to M while the "Only Question" dataset is used as shown in Table 5. So these two parameter values are used to find optimum seed number S. Because of having lower incorrectly classified instances while the "Question+Answer" dataser is used, optimum value for S is searched for "Question+Answer" dataset in 5,6,7,8,9,10,11,12,13,14,15 value set too. Best results are received with the highest and the most balanced precision, recall values and one of the lowest "InC. Classi." value while S is set to 15 on the "Only Question" dataset as shown in Table 6. Results show that best results of KMeans on GoDaddyDS are received by setting parameters as N:3 A:M I:500 S:15 and using the "Only Question" dataset.

1.2 EM Results

For the parameters of EM algorithm, default value is defined as 100 for maximum iteration number. On the tables, "InC. Classi." represents the incorrectly classified instance number depending on the Considered As section and used parameters are defined on the right side where "I:" represents the maximum iteration number, "N:" represents the number of clusters, "M:" represents the minimum standard deviation, "S:" represents the seed number. "NoClass" phrase under the "Considered as" section means, Weka did not attend any class name for that cluster, because the class label is attended before. But having clusters with same "ConsideredAs" class is acceptable, because of that the considered class number, which has most population in that cluster, is added after the "-" symbol for the "ConsideredAs" section of the table. Three datasets, which are made by using only question, only answer and both question and answer parts of the question-answer pairs, are used in EM tests. Used parameter values and results of the EM algorithm on datasets are given as subsections with the dataset names.

1.2.1 PrincetonDS

N is set to -1(algorithm sets the N by itself), 2, 3, 4, 5, 6, 7, 8 respectively while other parameters are set to default values. Then with the optimum N, M is set to 5.0E-7, 1.0E-6, 2.0E-6, 4.0E-6, 8.0E-6, 9.0E-6, 16.0E-6 respectively. And with the optimum cluster and minimum standard deviation parameters, S is set to 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 respectively.

Best results are received while N is set to 5 and the "Only Question" dataset is used as shown in Table 7. So this value is used to find optimum M. Best results are received while M is set to 8.0E-6 as shown in Table 8. Additionally, optimum parameter search for M is done while N is set to 6, because of having close results when N is set to 5 and 6. And better results are received while N is set to 5 if the results for N:5 are compared with the table the "Only Question for N:6". Optimum S is searched while N and M are set to the optimum values. Best results are received while S is set to 150 as shown in Table 9. Results show that best results of EM algorithm on PrincetonDS are received with highest and most balanced precision, recall and one of the lowest incorrectly classified values by setting parameters as N:5, M:8.0E-6, S:150 on Only Question dataset.

1.2.2 ParallelsDS

N is set to -1(algorithm sets the N by itself), 5, 6, 7, 8, 9, 10, 11 respectively while other parameters are set to default values. Then with the optimum N, M is set to 5.0E-7, 1.0E-6, 2.0E-6, 4.0E-6, 8.0E-6, 9.0E-6, 16.0E-6 respectively. And with the optimum cluster and minimum standard deviation parameters, S is set to 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 respectively.

The best results are received while N is set to 11 and the "Only Question" dataset is used as shown in Table 10. So this value is used to find optimum M. The best results are received while M is set to 1.0E-6 as shown in Table 11. Optimum S is searched while N and M are set to the optimum values. Best results are received while S is set to 100 as

shown in Table 12. Results show that best results of EM algorithm on ParallelDS are received with highest and most balanced precision, recall and one of the lowest incorrectly classified values by setting parameters as N:11, M: 1.0E-6, S:100 on Only Question dataset while.

1.2.3 GoDaddyDS

N is set to -1(algorithm sets the N by itself), 2, 3, 4, 5 respectively while other parameters are set to default values. Then with the optimum N, M is set to 5.0E-7, 1.0E-6, 2.0E-6, 4.0E-6, 8.0E-6, 9.0E-6, 16.0E-6 respectively. And with the optimum cluster and minimum standard deviation parameters, S is set to 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 respectively.

The best results are received while N is set to 3 and Only Question dataset is used as shown in Table 13. So this value is used to find optimum M. The best results are received while M is set to 2.0E-6 as shown in Table 14. Optimum S is searched while N and M are set to the optimum values. The best results are received while S is set to 100 as shown in Table 15. Results show that best results of EM algorithm on GoDaddyDS are received with highest and most balanced precision, recall and one of the lowest incorrectly classified values by setting parameters as N:3, M: 2.0E-6, S:100 on the “Only Question” dataset.

1.3 DBSCAN Results

DBSCAN algorithm is applied to the all three datasets with parameter values 0.9, 0.7, 0.2 for epsilon and 9, 7, 5, 2 for minimum points. But DBSCAN algorithm created only one cluster which consists of all instances while minimum points parameter is set to 2. As a result, DBSCAN is not able to cluster the instances of PrincetonDS, ParallelDS and GoDaddyDS.

1.4 Cosine Similarity Results

For the parameters of Cosine similarity measure (for more detail refer to Section 3 Methodology), default value isdefined as 0.05 for the Cosine Similarity threshold and 0.5 for the content value. For each dataset, content value is set to 0.3, 0.4, 0.5, 0.6, 0.7 values respectively and cosine similarity threshold is set to 0.03, 0.04, 0.05, 0.06 values respectively. Additionally, depending on the results, sometimes additional parameter values are used to see if it is possible to get better result. On the tables, "StandardD." represents the average standard deviation of the cluster, "Content" represents the content value parameter and "CosineSim.Thresh." represents the Cosine Similarity Threshold parameter value of Cosine Similarity method.

1.4.1 PrincetonDS

Content value parameter is set to 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the similarity threshold is set to default value. Then with the optimum content value, 0.03, 0.04, 0.05, 0.055, 0.06, 0.07 values are used to find the optimum similarity threshold parameter.

Best results for content value search are received while the content value is set to 0.4 as shown in Table 16. The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision and recall values with balanced cluster sizes while content value is set to 0.4 and Cosine similarity threshold is set to 0.04. Additionally, because of defining at least one cluster for each class label, results of Table 17 are considered as the best results on PrincetonDS.

1.4.2 ParallelsDS

Content value parameter is set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the similarity threshold is set to default value. Then with the optimum content value0.03,

0.04, 0.045, 0.05, 0.06 values are used to find the optimum similarity threshold parameter.

Best results for the content value search are received while the content value is set to 0.1 as shown in Table 18. The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision, recall values and balanced cluster sizes while content value is set to 0.1 and Cosine similarity threshold is set to 0.04 on ParallelsDS as showin Table 19.

1.4.3 GoDaddyDS

Content value parameter is set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the similarity threshold is set to default value. Then with the optimum content 0.03, 0.04, 0.05, 0.055, 0.06 values are used to find the optimum similarity threshold parameter.

Best results for the content value search are received while the content value is set to 0.2 as shown in Table 20 . The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision and recall values with balanced cluster sizes while content value is set to 0.2 and Cosine similarity threshold is set to 0.05 on GoDaddyDS as shown in Table 21.

1.5 Jaccard Distance Measure Results

For the parameter of Jaccard Distance measure(for more detail refer to Section 3 Methodology), default value is defined as 0.994 for the Jaccard distance threshold and 0.5 for the content value. For each dataset, content value is set to 0.3, 0.4, 0.5, 0.6, 0.7 respectively and Jaccard distance threshold is set to 0.992, 0.994, 0.996, 0.998 respectively. Additionally, depending on the results, sometimes additional parameter values are used to see if it is possible to get better results. On the tables, “StandardD.” represents the average standard deviation of the cluster, “Content” represents the content value parameter and “JaccardDist. Thresh.” represents the Jaccard distance threshold parameter value of Jaccard distance measure.

1.5.1 PrincetonDS

Content value parameter is set to 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 respectively while the distance threshold is set to default value. Then with the optimum content value 0.992, 0.994, 0.996, 0.997, 0.998, 0.9975 values are used to find the optimum Jaccard distance threshold parameter.

Best results for the content value search are received while the content value is set to 0.4 as shown in Table 22 . The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision and recall values with balanced cluster sizes while content value is set to 0.2 and Jaccard distance threshold is set to 0.9975 on PrincetonDS as shown in Table 23.

1.5.2 ParallelsDS

Content value parameter is set to 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the distance threshold is set to default value. Then with the optimum content value 0.992, 0.994, 0.996, 0.997, 0.998, 0.9975 values are used to find the optimum Jaccard distance threshold parameter.

Best results for the content value search are received while the content value is set to 0.4 as shown in Table 24. The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision and recall values with balanced cluster sizes while content value is set to 0.2 and Cosine similarity threshold is set to 0.997 on ParallelsDS. Additionally, because of results provide defining at least one cluster for each class label except label 7 while content value is 0.2 and distance threshold is 0.997 as shown in Table 25.

1.5.3 GoDaddyDS

Content value parameter is set to 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the distance threshold is set to default value. Then with the optimum content 0.992, 0.994, 0.996, 0.997, 0.998, 0.999, 0.9975 values are used to find the optimum Jaccard distance threshold parameter.

Best results for the content value search are received while the content value is set to 0.4 as shown in Table 26 . The final best results are received with one of the lowest Fwithin, one of the highest Fbetween and one of the highest precision and recall values with balanced cluster sizes while content value is set to 0.2 and Jaccard distance threshold is set to 0.9975 on GoDaddyDS as shown in Table 27.

1.6 CES Results

For the parameters of CES (for more detail refer to Section 3 Methodology), default value is defined as 0.5 for the content value (*content*), 0.006 for the tfidf threshold (*tfidf*), 0.3 for the creating core clusters threshold (*CC*), 0.003 for expanding clusters threshold(*E*) and 0.002 for sophisticating the clusters(*S*). Values for the parameters are defined firstly as "DefaultValues", which means *content*=0.5, *tfidf*=0.006, *CC*=0.3, *E*=0.003, *S*=0.002. If any parameter changed something else than the default values, it is mentioned such as "DefaultValues+content=0.4", which means content is set to 0.4 and all the other parameters are set to default values.

1.6.1 PrincetonDS

The *content* is set to 0.3, 0.4, 0.5, 0.6 respectively while the other parameters are set to default values. Then with the optimum content value, 0.003, 0.004, 0.0045, 0.005, 0.006, 0.007 values are used for the *tfidf* threshold parameter. With the optimum *content* and *tfidf* parameters, *CC* is set to 0.1, 0.2, 0.3, 0.4 respectively. With the optimum previous parameters, E is set to 0.001, 0.002, 0.003 and 0.004 respectively. Then finally with all previous optimum parameters, S is set to 0.009, 0.01, 0.002, 0.003 respectively to find the best results.

Results show that best result is received by setting parameters as *content*=0.4, *tfidf*=0.004, *CC*=0.2, *E*=0.002, *S*=0.001 as shown on Table 28.

1.6.2 ParallelsDS

The *content* is set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 respectively while the other parameters are set to default values. Then with the optimum content value, 0.003, 0.004, 0.0045, 0.005, 0.006, 0.007 values are used for the *tfidf* threshold parameter. With the

optimum *content* and *tfidf* parameters, *CC* is set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 respectively. With the optimum previous parameters, E is set to 0.001, 0.002, 0.003, 0.004, 0.005, 0.006 respectively. Then finally with all previous optimum parameters, S is set 0.0008, 0.0009, 0.001, 0.002, 0.003 respectively to find the best results.

Results show that best result is received by setting parameters as *content*=0.4, *tfidf*=0.004, *CC*=0.2, *E*=0.002, *S*=0.001 as shown on Table 29.

1.6.3 GoDaddyDS

The *content* is set to 0.3, 0.4, 0.5, 0.6 respectively while the other parameters are set to default values. Then with the optimum content value0.003, 0.004, 0.0045, 0.005, 0.006, 0.007 values are used for the *tfidf* threshold parameter. With the optimum *content* and *tfidf* parameters, *CC* is set to 0.1, 0.2, 0.3, 0.4 respectively. With the optimum previous parameters, E is set to 0.001, 0.002, 0.003 and 0.004 respectively. Then finally with all previous optimum parameters, S is set 0.0001, 0.0005, 0.0008, 0.0009, 0.001, 0.002, 0.003 respectively to find the best results.

Results show that best result is received by setting parameters as *content*=0.4, *tfidf*=0.004, *CC*=0.2, *E*=0.003, *S*=0.0009 as shown on Table 30.

1.7 CES+ Results

The same values for the parameters of the CES algorithm are used in the following , because both CES and CES+ algorithms give the same results with different time efficiency as shown in Table 31, Table 32 and Table 33. The tables in this section show that there is a big difference between time consumptions of CES and CES+.

1.8 TABLES

1.8.1 KMeans Results

1.8.1-1 PrincetonDS

Only Question					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	97.00	0.21	1.00	1.00	InC. Classi.: 77
Cluster2	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	95.00	0.21	1.00	1.00	InC. Classi.: 76
Cluster2	2.00	0.50	0.05	2.00	
Cluster3	1	1	0.05	5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	2.00	0.50	0.05	2.00	InC. Classi.: 72
Cluster2	86.00	0.23	1.00	3.00	
Cluster3	9.00	0.44	0.20	4.00	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	85.00	0.22	0.95	1.00	InC. Classi.: 72
Cluster2	1.00	0.50	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	9.00	0.44	0.20	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	82.00	0.22	0.90	1.00	InC. Classi.: 71
Cluster2	8.00	0.50	0.20	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	4.00	0.75	0.15	4.00	
Cluster5	2.00	0.50	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	81.00	0.22	0.90	1.00	InC. Classi.: 71
Cluster2	8.00	0.50	0.20	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	4.00	0.75	0.15	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	No Class-5	
Cluster7	1.00	0.50	0.05	No Class-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10

Cluster1	79.00	0.23	0.90	1.00	InC. Classi.: 71
Cluster2	8.00	0.50	0.20	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	4.00	0.75	0.15	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	2.00	0.50	0.05	NoClass-2	
Cluster8	2.00	1.00	20-Feb	NoClass-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	97.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	90.00	0.22	1.00	1.00	InC. Classi.: 73
Cluster2	7.00	0.57	0.20	2.00	
Cluster3	1.00	1.00	20-Jan	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	71.00	0.25	0.90	1.00	InC. Classi.: 67
Cluster2	7.00	0.57	0.20	2.00	
Cluster3	19.00	0.42	0.40	3.00	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	71.00	0.27	0.95	1.00	InC. Classi.: 64
Cluster2	7.00	0.57	0.20	2.00	
Cluster3	18.00	0.56	0.50	3.00	
Cluster4	1.00	1.00	0.05	NoClass-3	
Cluster5	1.00	1.00	0.05	5.00	

Table 1: Best results for the optimum N of KMeans on PrincetonDS

Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	67.00	0.27	0.90	1.00	InC. Classi.: 65
Cluster2	5.00	0.40	0.10	2.00	
Cluster3	18.00	0.56	0.50	3.00	
Cluster4	1.00	1.00	0.05	3.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-3	

Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	66.00	0.27	0.90	1.00	InC. Classi.: 65
Cluster2	5.00	0.40	0.10	2.00	
Cluster3	18.00	0.56	0.50	3.00	
Cluster4	6.00	0.33	0.10	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-3	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	66.00	0.26	0.85	1.00	InC. Classi.: 70
Cluster2	3.00	0.67	0.10	2.00	
Cluster3	15.00	0.40	0.30	3.00	
Cluster4	6.00	0.33	0.10	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	5.00	0.40	0.10	NoClass-2	
Cluster8	1.00	1.00	0.05	NoClass-3	

Search for optimum S while N is set to 5

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:5
Cluster1	3.00	1.00	0.15	1.00	InC. Classi.: 71
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	91.00	0.22	1.00	3.00	
Cluster4	2.00	1.00	0.10	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:6
Cluster1	80.00	0.24	0.95	1.00	InC. Classi.:67
Cluster2	4.00	0.75	0.15	2.00	
Cluster3	5.00	1.00	0.25	3.00	
Cluster4	1.00	1.00	0.05	NoClass-3	
Cluster5	8.00	0.50	0.20	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:7
Cluster1	71.00	0.27	0.95	1.00	InC. Classi.:57
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	11.00	0.91	0.50	3.00	
Cluster4	9.00	0.56	0.25	4.00	
Cluster5	5.00	1.00	0.25	5.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:8
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:73
Cluster2	7.00	0.57	0.20	2.00	
Cluster3	89.00	0.21	0.95	3.00	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:9
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:74
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	93.00	0.22	1.00	3.00	
Cluster4	1.00	1.00	0.05	4.00	
Cluster5	1.00	1.00	0.05	1.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	71.00	0.27	0.95	1.00	InC. Classi.: 64
Cluster2	7.00	0.57	0.20	2.00	
Cluster3	18.00	0.56	0.50	3.00	
Cluster4	1.00	1.00	0.05	NoClass-3	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:11
Cluster1	90.00	0.22	1.00	1.00	InC. Classi.: 72
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	5.00	0.80	0.20	3.00	
Cluster4	1.00	1.00	0.05	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:12
Cluster1	9.00	0.78	0.35	1.00	InC. Classi.:65
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	78.00	0.24	0.95	3.00	
Cluster4	9.00	0.56	0.25	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:13
Cluster1	8.00	0.75	0.30	1.00	InC. Classi.:61
Cluster2	9.00	0.78	0.35	2.00	
Cluster3	64.00	0.25	0.80	3.00	
Cluster4	9.00	0.56	0.25	4.00	
Cluster5	8.00	0.38	0.15	5.00	

Table 2: Best results on PrincetonDS with the optimum parameter values of KMeans

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:14
Cluster1	1.00	1.00	0.05	NoClass-2	InC. Classi.:69
Cluster2	6.00	0.50	0.15	2.00	
Cluster3	85.00	0.24	1.00	3.00	
Cluster4	1.00	1.00	0.05	4.00	
Cluster5	5.00	1.00	0.25	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:15
Cluster1	4.00	0.25	0.05	1.00	InC. Classi.:67
Cluster2	79.00	0.24	0.95	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	9.00	0.56	0.25	4.00	
Cluster5	5.00	1.00	0.25	5.00	

Only Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	97.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	96.00	0.21	2.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	1.00	1.00	20-Jan	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	1.00	1.00	0.05	NoClass-5	InC. Classi.:75
Cluster2	10.00	0.30	0.15	2.00	
Cluster3	86.00	0.22	0.95	4.00	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:76
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	93.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	

Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	92.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	91.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Cluster8	1.00	1.00	20-Jan	NoClass-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	97.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	96.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	1.00	1.00	39101.00	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:M I:500 S:10
Cluster1	71.00	0.25	0.90	1.00	InC. Classi.:68
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	25.00	0.44	0.55	3.00	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:76
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	93.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	92.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	91.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Cluster8	1.00	1.00	0.05	NoClass-5	

Question+Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:75
Cluster2	4.00	0.75	0.15	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	93.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	4.00	0.75	20-Mar	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	95.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	5.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:76
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	93.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	92.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	91.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Cluster8	1.00	1.00	0.05	NoClass-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	76.00	0.24	0.90	1.00	InC. Classi.:70
Cluster2	22.00	0.45	0.50	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	75.00	0.24	0.90	1.00	InC. Classi.:68
Cluster2	22.00	0.50	0.55	3.00	
Cluster3	1.00	1.00	0.05	5.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:M I:500 S:10
Cluster1	95.00	0.21	1.00	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	NoClass-5	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:76
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	NoClass-5	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	93.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	92.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	91.00	0.22	1.00	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Cluster8	1.00	1.00	0.05	NoClass-2	

1.8.1-2 ParallelDS

Only Question					
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	84.00	0.29	0.92	1.00	InC. Classi.:62
Cluster2	1.00	1.00	0.04	NoClass-1	
Cluster3	13.00	1.00	0.87	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	82.00	0.29	0.92	1.00	InC. Classi.:61
Cluster2	2.00	0.50	0.05	2.00	
Cluster3	13.00	1.00	0.87	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	81.00	0.30	0.92	1.00	InC. Classi.:60
Cluster2	2.00	0.50	0.05	2.00	
Cluster3	13.00	1.00	0.87	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	18.00	1.00	0.69	1.00	InC. Classi.:48
Cluster2	64.00	0.30	0.95	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	

Table 3: Best results for the optimum N of KMEans on PrincetonDS

Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:E I:500 S:10
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:51
Cluster2	64.00	0.30	0.95	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	3.00	1.00	0.12	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:E I:500 S:10
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:51
Cluster2	64.00	0.30	0.95	2.00	
Cluster3	12.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	2.00	1.00	26-Feb	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:51
Cluster2	63.00	0.30	0.95	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	11.00	0.13	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	2.00	1.00	0.08	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	88.00	0.27	0.92	1.00	InC. Classi.:66
Cluster2	1.00	1.00	0.04	NoClas-1	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClas-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	86.00	0.28	0.96	1.00	InC. Classi.:65
Cluster2	1.00	0.50	0.05	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	85.00	0.28	0.92	1.00	InC. Classi.:64
Cluster2	2.00	0.50	0.05	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.08	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	83.00	0.27	0.85	1.00	InC. Classi.:66
Cluster2	2.00	0.50	0.05	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	2.00	1.00	0.08	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:M I:500 S:10
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:66
Cluster2	80.00	0.24	0.95	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	2.00	1.00	0.08	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:M I:500 S:10
Cluster1	14.00	1.00	0.54	1.00	InC. Classi.:55
Cluster2	68.00	0.28	0.95	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	0.50	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	2.00	1.00	26-Feb	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	14.00	1.00	0.54	1.00	InC. Classi.:55
Cluster2	67.00	0.28	0.95	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.13	NoClass-5	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	2.00	1.00	0.08	NoClass-1	

Search for optimum S while N is set to 8

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:5
Cluster1	28.00	0.57	0.62	1.00	InC. Classi.:57
Cluster2	50.00	0.24	0.60	2.00	
Cluster3	13.00	1.00	0.87	3.00	
Cluster4	3.00	1.00	0.12	NoClass-1	
Cluster5	2.00	1.00	0.13	NoClass-3	
Cluster6	1.00	1.00	0.05	NoClass-2	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	0.50	0.20	8.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:6
Cluster1	6.00	1.00	0.23	1.00	InC. Classi.:70
Cluster2	73.00	0.23	0.85	2.00	
Cluster3	13.00	0.31	0.27	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	0.50	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	2.00	1.00	0.08	NoClass-1	
Cluster8	1.00	0.50	0.20	8.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:7
Cluster1	17.00	1.00	0.65	1.00	InC. Classi.:55
Cluster2	68.00	0.28	0.95	2.00	
Cluster3	1.00	1.00	0.08	NoClass-4	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	9.00	0.89	1.00	5.00	
Cluster6	2.00	0.50	0.08	NoClass-4	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:8
Cluster1	16.00	1.00	0.62	1.00	InC. Classi.:54
Cluster2	55.00	0.29	0.80	2.00	
Cluster3	3.00	1.00	0.25	NoClass-4	
Cluster4	12.00	0.75	0.75	4.00	
Cluster5	3.00	0.33	0.13	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	3.00	0.67	0.10	NoClass-2	
Cluster8	7.00	0.43	0.60	8.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:9
Cluster1	90.00	0.26	0.88	1.00	InC. Classi.:71
Cluster2	2.00	1.00	0.08	NoClass-1	
Cluster3	3.00	1.00	0.20	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.07	NoClass-3	

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	18.00	1.00	0.69	1.00	InC. Classi.:48
Cluster2	64.00	0.30	0.95	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	

Table 4: Best results on ParallelsDS with the optimum parameter values of KMeans

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:11
Cluster1	68.00	0.35	0.92	1.00	InC. Classi.:53
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	2.00	1.00	0.08	NoClass-1	
Cluster4	12.00	1.00	1.00	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	0.50	0.14	6.00	
Cluster7	12.00	0.58	1.00	7.00	
Cluster8	1.00	0.50	0.20	8.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:12
Cluster1	72.00	0.36	1.00	1.00	InC. Classi.:50
Cluster2	1.00	1.00	0.08	NoClass-4	
Cluster3	14.00	1.00	0.93	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	9.00	0.89	1.00	5.00	
Cluster6	1.00	1.00	0.14	NoClass-7	
Cluster7	1.00	1.00	0.14	NoClass-7	
Cluster8	1.00	1.00	0.07	NoClass-3	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:13
Cluster1	76.00	0.32	0.92	1.00	InC. Classi.:56
Cluster2	1.00	1.00	0.04	NoClass-1	
Cluster3	13.00	1.00	0.87	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	6.00	0.67	0.57	7.00	
Cluster8	1.00	1.00	0.20	8.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:14
Cluster1	72.00	0.36	1.00	1.00	InC. Classi.:58
Cluster2	20.00	0.55	0.55	2.00	
Cluster3	2.00	1.00	0.13	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.05	NoClass-2	
Cluster6	1.00	1.00	0.05	NoClass-2	
Cluster7	1.00	1.00	0.05	NoClass-2	
Cluster8	2.00	1.00	0.40	8.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:15
Cluster1	72.00	0.32	0.88	1.00	InC. Classi.:59
Cluster2	3.00	0.33	0.05	2.00	
Cluster3	1.00	1.00	0.08	NoClass-4	
Cluster4	3.00	1.00	0.25	4.00	
Cluster5	6.00	0.83	0.63	5.00	
Cluster6	12.00	0.58	1.00	6.00	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	2.00	0.50	0.20	8.00	

Only Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	8.00	1.00	0.31	1.00	InC. Classi.:57
Cluster2	75.00	0.27	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	6.00	1.00	0.23	1.00	InC. Classi.:71
Cluster2	19.00	0.32	0.23	2.00	
Cluster3	71.00	0.21	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	2.00	1.00	0.29	6.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	8.00	1.00	0.31	1.00	InC. Classi.:54
Cluster2	72.00	0.28	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	2.00	1.00	0.29	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	9.00	0.56	0.19	1.00	InC. Classi.:76
Cluster2	83.00	0.20	0.85	2.00	
Cluster3	3.00	1.00	0.12	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:E I:500 S:10
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:76
Cluster2	9.00	0.44	0.20	2.00	
Cluster3	82.00	0.18	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:E I:500 S:10
Cluster1	7.00	1.00	0.27	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	73.00	0.21	1.00	3.00	
Cluster4	11.00	0.18	0.17	4.00	
Cluster5	1.00	1.00	0.05	NoClass-2	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	2.00	0.50	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.05	NoClass-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:76
Cluster2	3.00	0.67	0.10	2.00	
Cluster3	80.00	0.19	1.00	3.00	
Cluster4	2.00	0.50	0.08	4.00	
Cluster5	2.00	0.50	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	2.00	0.50	0.14	7.00	
Cluster8	1.00	1.00	0.05	NoClass-2	
Cluster9	2.00	0.50	0.05	NoClass-2	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:61
Cluster2	79.00	0.25	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	0.04	0.04	NoClass-1	
Cluster5	1.00	0.04	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:55
Cluster2	73.00	0.27	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	6.00	1.00	0.86	6.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:54
Cluster2	72.00	0.28	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-4	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	6.00	1.00	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:54
Cluster2	71.00	0.28	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	6.00	1.00	0.86	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:M I:500 S:10
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:54
Cluster2	70.00	0.29	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	6.00	1.00	0.86	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:M I:500 S:10
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:55
Cluster2	70.00	0.29	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	6.00	1.00	0.86	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:49
Cluster2	63.00	0.32	1.00	2.00	
Cluster3	15.00	1.00	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	7.00	1.00	0.88	5.00	
Cluster6	6.00	1.00	0.86	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-5	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	

Question+Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:77
Cluster2	94.00	0.21	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	2.00	1.00	0.08	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:77
Cluster2	91.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	3.00	0.33	0.08	4.00	
Cluster5	1.00	1.00	0.08	NoClass-4	
Cluster6	2.00	1.00	0.08	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:76
Cluster2	90.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	3.00	0.33	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.08	NoClass-4	
Cluster7	2.00	1.00	0.08	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:76
Cluster2	89.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-3	
Cluster4	3.00	0.33	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.08	NoClass-4	
Cluster7	2.00	1.00	0.08	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:76
Cluster2	88.00	0.45	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	3.00	0.33	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.08	NoClass-4	
Cluster7	2.00	1.00	0.08	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:E I:500 S:10
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:76
Cluster2	88.00	0.23	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	3.00	0.33	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.08	NoClass-4	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	3.00	0.67	0.08	1.00	InC. Classi.:76
Cluster2	88.00	0.23	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:74
Cluster2	92.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:6 A:M I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:74
Cluster2	91.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:7 A:M I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:73
Cluster2	90.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:8 A:M I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:73
Cluster2	89.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:9 A:M I:500 S:10
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:73
Cluster2	88.00	0.23	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:10 A:M I:500 S:10
Cluster1	14.00	0.71	0.38	1.00	InC. Classi.:70
Cluster2	78.00	0.23	0.90	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:11 A:E I:500 S:10
Cluster1	18.00	0.50	0.35	1.00	InC. Classi.:74
Cluster2	73.00	0.21	0.75	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	

1.8.1-1 GoDaddyDS

Only Questions					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	97.00	0.39	0.93	1.00	InC. Classi.: 62
Cluster2	3.00	1.00	0.07	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	3.00	1.00	0.07	1.00	InC. Classi.: 63
Cluster2	95.00	0.37	1.00	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	3.00	1.00	0.07	1.00	InC. Classi.: 63
Cluster2	94.00	0.36	1.00	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	93.00	0.38	0.85	1.00	InC. Classi.: 64
Cluster2	1.00	1.00	0.03	2.00	
Cluster3	3.00	1.00	0.07	NoClass-1	
Cluster4	2.00	1.00	0.05	NoClass-1	
Cluster5	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	4.00	1.00	0.10	1.00	InC. Classi.: 62
Cluster2	96.00	0.35	1.00	2.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	4.00	1.00	0.10	1.00	InC. Classi.:62
Cluster2	94.00	0.36	1.00	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	

Table 5: Best results for the optimum N of KMeans on GoDaddyDS

Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:M I:500 S:10
Cluster1	4.00	1.00	0.10	1.00	InC. Classi.: 62
Cluster2	93.00	0.37	1.00	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	4.00	1.00	0.10	1.00	InC. Classi.: 63
Cluster2	92.00	0.36	0.97	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Cluster5	1.00	1.00	0.03	NoClass-2	

Search for optimum S while N is set to 3

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:5
Cluster1	18.00	1.00	0.44	1.00	InC. Classi.:50
Cluster2	72.00	0.42	0.88	2.00	
Cluster3	10.00	0.20	0.08	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:6
Cluster1	98.00	0.40	0.95	1.00	InC. Classi.:61
Cluster2	1.00	1.00	0.02	NoClass-1	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:7
Cluster1	85.00	0.44	0.90	1.00	InC. Classi.:53
Cluster2	9.00	1.00	0.26	2.00	
Cluster3	6.00	0.17	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:8
Cluster1	19.00	1.00	0.46	1.00	InC. Classi.:47
Cluster2	78.00	0.24	0.56	2.00	
Cluster3	3.00	1.00	0.07	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:9
Cluster1	74.00	0.50	0.90	1.00	InC. Classi.:41
Cluster2	23.00	0.96	0.65	2.00	
Cluster3	3.00	1.00	0.07	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	4.00	1.00	0.10	1.00	InC. Classi.:62
Cluster2	94.00	0.36	1.00	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:11
Cluster1	19.00	1.00	0.46	1.00	InC. Classi.:30
Cluster2	64.00	0.53	1.00	2.00	
Cluster3	17.00	1.00	0.68	3.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:12
Cluster1	64.00	0.53	1.00	1.00	InC. Classi.:37
Cluster2	35.00	0.80	0.82	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:13
Cluster1	87.00	0.43	0.90	1.00	InC. Classi.:54
Cluster2	12.00	0.67	0.24	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:14
Cluster1	96.00	0.39	0.90	1.00	InC. Classi.:63
Cluster2	3.00	1.00	0.07	NoClass-1	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:15
Cluster1	20.00	1.00	0.49	1.00	InC. Classi.:25
Cluster2	33.00	0.91	0.97	2.00	
Cluster3	47.00	0.53	1.00	3.00	

Table 6: Best results on GoDaddyDS with the optimum parameter values of KMeans

Only Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.:56
Cluster2	3.00	1.00	0.09	2.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	95.00	0.41	0.95	1.00	InC. Classi.: 58
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	94.00	0.40	0.93	1.00	InC. Classi.: 59
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	78.00	0.44	0.83	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	16.00	0.38	0.24	3.00	
Cluster4	2.00	1.00	0.05	NoClass-1	
Cluster5	1.00	1.00	0.02	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.: 56
Cluster2	3.00	1.00	0.09	2.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:M I:500 S:10
Cluster1	95.00	0.41	0.95	1.00	InC. Classi.: 58
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	59.00	0.46	0.66	1.00	InC. Classi.: 58
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	36.00	0.33	0.48	3.00	
Cluster4	1.00	1.00	0.02	NoClass-1	
Cluster5	1.00	1.00	0.05	NoClass-1	

Search for optimum S while N is set to 3

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:5
Cluster1	97.00	0.40	0.95	1.00	InC. Classi.: 60
Cluster2	2.00	1.00	0.05	NoClass-1	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:6
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:7
Cluster1	5.00	0.80	0.10	1.00	InC. Classi.: 62
Cluster2	94.00	0.36	1.00	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:8
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:9
Cluster1	98.00	0.42	1.00	1.00	InC. Classi.: 57
Cluster2	1.00	1.00	0.03	2.00	
Cluster3	1.00	1.00	0.03	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:11
Cluster1	98.00	0.41	0.98	1.00	InC. Classi.: 59
Cluster2	1.00	1.00	0.02	NoClass-1	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:12
Cluster1	92.00	0.45	1.00	1.00	InC. Classi.: 51
Cluster2	7.00	1.00	0.21	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:13
Cluster1	83.00	0.49	1.00	1.00	InC. Classi.: 43
Cluster2	1.00	1.00	0.04	NoClass-3	
Cluster3	16.00	1.00	0.64	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:14
Cluster1	58.00	0.47	0.66	1.00	InC. Classi.: 59
Cluster2	39.00	0.36	0.41	2.00	
Cluster3	3.00	1.00	0.09	NoClass-2	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:15
Cluster1	82.00	0.49	0.98	1.00	InC. Classi.: 43
Cluster2	1.00	1.00	0.02	NoClass-1	
Cluster3	17.00	1.00	0.68	3.00	

Question+Answer					
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:E I:500 S:10
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.:56
Cluster2	3.00	1.00	0.09	2.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:E I:500 S:10
Cluster1	95.00	0.41	0.95	1.00	InC. Classi.: 58
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	

Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:E I:500 S:10
Cluster1	94.00	0.40	0.93	1.00	InC. Classi.: 59
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:E I:500 S:10
Cluster1	86.00	0.42	0.88	1.00	InC. Classi.: 56
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	8.00	0.63	0.20	3.00	
Cluster4	2.00	1.00	0.05	NoClass-1	
Cluster5	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:2 A:M I:500 S:10
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.: 56
Cluster2	3.00	1.00	0.09	2.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:4 A:M I:500 S:10
Cluster1	94.00	0.40	0.93	1.00	InC. Classi.: 59
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:5 A:M I:500 S:10
Cluster1	55.00	0.45	0.61	1.00	InC. Classi.: 60
Cluster2	40.00	0.38	0.44	2.00	
Cluster3	3.00	1.00	0.09	NoClass-2	
Cluster4	1.00	1.00	0.02	NoClass-1	
Cluster5	1.00	1.00	0.02	NoClass-1	

Search for optimum S while N is set to 3

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:5
Cluster1	97.00	0.40	0.95	1.00	InC. Classi.: 60
Cluster2	2.00	1.00	0.05	NoClass-1	
Cluster3	1.00	1.00	0.04	3.00	

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:6
Cluster1	97.00	0.39	0.93	1.00	InC. Classi.: 62
Cluster2	1.00	1.00	0.02	NoClass-1	
Cluster3	2.00	1.00	0.05	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:7
Cluster1	5.00	1.00	0.12	1.00	InC. Classi.: 61
Cluster2	94.00	0.36	0.83	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:8
Cluster1	7.00	1.00	0.17	1.00	InC. Classi.: 59
Cluster2	92.00	0.37	1.00	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:9
Cluster1	98.00	0.42	1.00	1.00	InC. Classi.: 57
Cluster2	1.00	1.00	0.03	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:10
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:11
Cluster1	98.00	0.41	0.98	1.00	InC. Classi.: 59
Cluster2	1.00	1.00	0.02	NoClass-1	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:12
Cluster1	96.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:13
Cluster1	82.00	0.50	1.00	1.00	InC. Classi.: 42
Cluster2	1.00	1.00	0.04	NoClass-3	
Cluster3	17.00	1.00	0.68	3.00	
Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:14
Cluster1	23.00	0.96	0.54	1.00	InC. Classi.: 48
Cluster2	74.00	0.41	0.88	2.00	
Cluster3	3.00	1.00	0.09	NoClass-2	

Kmeans	Size	Precision	Recall	ConsideredAs	N:3 A:M I:500 S:15
Cluster1	96.00	0.42	0.98	1.00	InC. Classi.: 57
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.02	NoClass-1	

1.8.2 EM Results

1.8.2-1 PrincetonDS

Only Question					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.: 62
Cluster2	43.00	0.35	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	17.00	0.41	0.35	4.00	
Cluster5	33.00	0.33	0.55	5.00	
CLuster6	1.00	1.00	0.05	NoClass-1	
Cluster7	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	12.00	0.92	0.55	1.00	InC. Classi.: 67
Cluster2	86.00	0.23	1.00	2.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	19.00	0.37	0.35	1.00	InC. Classi.:73
Cluster2	77.00	0.23	0.90	4.00	
Cluster3	2.00	0.50	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:66
Cluster2	32.00	0.41	0.65	3.00	
Cluster3	30.00	0.27	0.40	4.00	
Cluster4	35.00	0.29	0.50	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:59
Cluster2	39.00	0.38	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	41.00	0.34	0.70	4.00	
Cluster5	16.00	0.50	0.40	5.00	

Table 7: Best results for the optimum N of EM on PrincetonDS

EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:56
Cluster2	46.00	0.37	0.85	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	35.00	0.43	0.75	4.00	
Cluster5	13.00	0.54	0.35	5.00	
Cluster6	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:62
Cluster2	43.00	0.35	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	17.00	0.41	0.35	4.00	
Cluster5	33.00	0.33	0.55	5.00	
Cluster6	1.00	1.00	0.05	1.00	
Cluster7	1.00	1.00	0.05	1.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	12.00	0.42	0.25	1.00	InC. Classi.:59
Cluster2	17.00	0.53	0.45	2.00	
Cluster3	29.00	0.41	0.60	3.00	
Cluster4	7.00	0.57	0.20	4.00	
Cluster5	30.00	0.30	0.45	5.00	
Cluster6	1.00	1.00	0.05	NoClass-1	
Cluster7	1.00	1.00	0.05	NoClass-3	
Cluster8	1.00	1.00	0.05	NoClass-1	

Search for optimum M while N is set to 5

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:2.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:64
Cluster2	28.00	0.32	0.45	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	33.00	0.36	0.60	4.00	
Cluster5	35.00	0.31	0.55	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:59
Cluster2	39.00	0.38	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	41.00	0.34	0.70	4.00	
Cluster5	16.00	0.50	0.40	5.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:5.0E-7 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:59
Cluster2	39.00	0.38	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	41.00	0.34	0.70	4.00	
Cluster5	16.00	0.50	0.40	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:25.0E-8 S:100
Cluster1	11.00	0.64	0.35	1.00	InC. Classi.:60
Cluster2	60.00	0.30	0.90	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	25.00	0.48	0.60	4.00	
Cluster5	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:4.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:59
Cluster2	42.00	0.36	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	38.00	0.34	0.65	4.00	
Cluster5	15.00	0.60	0.45	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:55
Cluster2	41.00	0.37	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	33.00	0.42	0.70	4.00	
Cluster5	21.00	0.52	0.55	5.00	

Table 8: Best results for the optimum M and N paremeters of EM on PrincetonDS

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:16.0E-6 S:100
Cluster1	2.00	1.00	0.05	1.00	InC. Classi.:67
Cluster2	24.00	0.38	0.45	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	63.00	0.24	0.75	4.00	
Cluster5	8.00	0.50	0.20	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:9.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:66
Cluster2	15.00	0.40	0.30	2.00	
Cluster3	46.00	0.26	0.60	3.00	
Cluster4	34.00	0.35	0.60	4.00	
Cluster5	1.00	1.00	0.05	NoClass-3	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-5 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:67
Cluster2	24.00	0.38	0.45	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	63.00	0.24	0.75	4.00	
Cluster5	8.00	0.50	0.20	5.00	

Search for optimum S while M is set to 8.0E-6 and N is set to 5

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:50
Cluster1	31.00	0.42	0.65	1.00	InC. Classi.:60
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	31.00	0.45	0.70	3.00	
Cluster4	34.00	0.29	0.50	4.00	
Cluster5	1.00	1.00	0.05	NoClass-3	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:60
Cluster1	13.00	0.69	0.45	1.00	InC. Classi.:62
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	43.00	0.33	0.70	3.00	
Cluster4	34.00	0.32	0.55	4.00	
Cluster5	7.00	0.14	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:70
Cluster1	8.00	0.38	0.15	1.00	InC. Classi.:65
Cluster2	1.00	1.00	0.05	NoClass-1	
Cluster3	28.00	0.32	0.45	3.00	
Cluster4	36.00	0.39	0.70	4.00	
Cluster5	25.00	0.28	0.35	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:80
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	94.00	0.21	1.00	4.00	
Cluster5	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:90
Cluster1	39.00	0.28	0.55	1.00	InC. Classi.:66
Cluster2	7.00	0.86	0.30	2.00	
Cluster3	1.00	1.00	0.05	NoClass-1	
Cluster4	50.00	0.30	0.75	4.00	
Cluster5	1.00	1.00	0.05	NoClass-4	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:55
Cluster2	41.00	0.37	0.75	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	33.00	0.42	0.70	4.00	
Cluster5	21.00	0.52	0.55	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:110
Cluster1	67.00	0.30	1.00	1.00	InC. Classi.:57
Cluster2	15.00	0.67	0.50	2.00	
Cluster3	3.00	0.33	0.05	3.00	
Cluster4	12.00	0.83	0.50	4.00	
Cluster5	1.00	1.00	0.05	NoClass-4	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:120
Cluster1	18.00	0.50	0.45	1.00	InC. Classi.:69
Cluster2	77.00	0.25	0.95	2.00	
Cluster3	1.00	1.00	0.05	NoClass-1	
Cluster4	1.00	1.00	0.05	NoClass-1	
Cluster5	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:130
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:65
Cluster2	48.00	0.29	0.70	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	4.00	0.75	0.15	4.00	
Cluster5	44.00	0.32	0.70	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:140
Cluster1	77.00	0.26	1.00	1.00	InC. Classi.:64
Cluster2	18.00	0.67	0.60	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-2	
Cluster5	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:150
Cluster1	24.00	0.42	0.50	1.00	InC. Classi.:57
Cluster2	34.00	0.44	0.75	2.00	
Cluster3	34.00	0.29	0.50	3.00	
Cluster4	5.00	1.00	0.25	4.00	
Cluster5	1.00	1.00	0.05	5.00	

Table 9: Best results with the optimum parameters of EM on PrincetonDS

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:8.0E-6 S:160
Cluster1	9.00	0.67	0.30	1.00	InC. Classi.:60
Cluster2	8.00	0.88	0.35	2.00	
Cluster3	28.00	0.25	0.35	3.00	
Cluster4	4.00	1.00	0.20	4.00	
Cluster5	49.00	0.29	0.70	5.00	

Search for M on Only Question dataset for N:6					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:2.0E-6 S:100
Cluster1	3.00	1.00	0.05	1.00	InC. Classi.:65
Cluster2	28.00	0.29	0.40	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	21.00	0.33	0.35	4.00	
Cluster5	44.00	0.32	0.70	5.00	
Cluster6	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:5.0E-7 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:61
Cluster2	38.00	0.37	0.70	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	40.00	0.35	0.70	4.00	
Cluster5	17.00	0.41	0.35	5.00	
Cluster6	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:25.0E-8 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:61
Cluster2	17.00	0.53	0.45	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	25.00	0.44	0.55	4.00	
Cluster5	53.00	0.28	0.75	5.00	
Cluster6	1.00	1.00	0.05	NoClas-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:4.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:61
Cluster2	53.00	0.34	0.90	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	29.00	0.31	0.45	4.00	
Cluster5	12.00	0.58	0.35	5.00	
Cluster6	1.00	1.00	0.05	NoClass-1	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:6M:1.6.0E-5 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:60
Cluster2	59.00	0.34	1.00	2.00	
Cluster3	7.00	0.71	0.25	3.00	
Cluster4	28.00	0.39	0.55	4.00	
Cluster5	1.00	1.00	0.05	NoClass-3	
Cluster6	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:9.0E-6 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:63
Cluster2	59.00	0.34	1.00	2.00	
Cluster3	7.00	0.71	0.25	3.00	
Cluster4	28.00	0.39	0.55	4.00	
Cluster5	1.00	1.00	0.05	NoClass-3	
Cluster6	1.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-5 S:100
Cluster1	2.00	1.00	0.10	1.00	InC. Classi.:62
Cluster2	53.00	0.34	0.90	2.00	
Cluster3	7.00	0.71	0.25	3.00	
Cluster4	34.00	0.32	0.55	4.00	
Cluster5	1.00	1.00	0.05	NoClass-3	
Cluster6	1.00	1.00	0.05	NoClass-1	

OnlyAnswer					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	7.00	0.43	0.15	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	87.00	0.23	1.00	3.00	
Cluster4	1.00	1.00	0.05	NoClass-2	
Cluster5	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	96.00	0.21	1.00	1.00	InC. Classi.: 76
Cluster2	2.00	1.00	0.10	2.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	7.00	0.43	0.15	1.00	InC. Classi.:74
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	90.00	0.22	1.00	3.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	6.00	0.50	0.15	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	89.00	0.22	1.00	4.00	
Cluster4	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	7.00	0.43	0.15	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	87.00	0.23	1.00	3.00	
Cluster4	1.00	1.00	0.05	NoClass-2	
Cluster5	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	7.00	0.43	0.15	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	3.00	0.33	0.05	3.00	
Cluster4	84.00	0.23	0.95	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	7.00	0.43	0.15	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	3.00	0.33	0.05	3.00	
Cluster4	83.00	0.23	0.95	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-2	
Cluster7	1.00	1.00	0.05	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:65
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	76.00	0.25	0.95	3.00	
Cluster4	15.00	0.67	0.50	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-4	
Cluster7	1.00	1.00	0.05	NoClass-3	
Cluster8	1.00	1.00	0.05	NoClass-2	

Question+Answer					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	98.00	0.20	1.00	1.00	InC. Classi.:78
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	96.00	0.21	1.00	1.00	InC. Classi.: 76
Cluster2	2.00	1.00	0.10	2.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	95.00	0.21	1.00	1.00	InC. Classi.:75
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	94.00	0.21	1.00	1.00	InC. Classi.:75
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	NoClass-2	
Cluster4	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	7.00	0.29	0.10	1.00	InC. Classi.:73
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	87.00	0.22	0.95	4.00	
Cluster5	1.00	1.00	0.05	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	92.00	0.22	1.00	1.00	InC. Classi.:74
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	1.00	1.00	0.05	NoClass-5	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:73
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	3.00	
Cluster4	91.00	0.22	1.00	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	1.00	1.00	0.05	1.00	InC. Classi.:74
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	90.00	0.21	0.95	3.00	
Cluster4	1.00	1.00	0.05	4.00	
Cluster5	1.00	1.00	0.05	5.00	
Cluster6	1.00	1.00	0.05	NoClass-5	
Cluster7	1.00	1.00	0.05	NoClass-5	
Cluster8	1.00	1.00	0.05	NoClass-3	

1.8.2-2 ParallelsDS

Only Question					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	8.00	0.88	0.27	1.00	InC. Classi.:64
Cluster2	46.00	0.28	0.65	2.00	
Cluster3	19.00	0.47	0.60	3.00	
Cluster4	18.00	0.22	0.57	6.00	
Cluster5	9.00	0.33	0.43	7.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:65
Cluster2	1.00	1.00	0.04	NoClass-1	
Cluster3	56.00	0.27	1.00	3.00	
Cluster4	22.00	0.36	0.67	4.00	
Cluster5	2.00	1.00	0.25	5.00	
Cluster6	14.00	0.36	0.71	6.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	12.00	1.00	0.46	1.00	InC. Classi.:57
Cluster2	29.00	0.41	0.60	2.00	
Cluster3	18.00	0.50	0.60	3.00	
Cluster4	1.00	1.00	0.14	4.00	
Cluster5	11.00	0.18	0.25	5.00	
Cluster6	27.00	0.26	1.00	6.00	
Cluster7	2.00	1.00	0.29	7.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	28.00	0.61	0.65	1.00	InC. Classi.:49
Cluster2	12.00	0.58	0.35	2.00	
Cluster3	14.00	0.64	0.60	3.00	
Cluster4	18.00	0.50	0.75	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	12.00	0.42	0.71	6.00	
Cluster7	1.00	1.00	0.08	NoClass-4	
Cluster8	14.00	0.29	0.80	8.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:9 M:1.0E-6 S:100
Cluster1	26.00	0.73	0.73	1.00	InC. Classi.:52
Cluster2	19.00	0.42	0.40	2.00	
Cluster3	30.00	0.47	0.93	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	1.00	1.00	0.14	NoClass-7	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	20.00	0.25	0.71	7.00	
Cluster8	1.00	1.00	0.05	NoCLass-2	
Cluster9	1.00	1.00	0.07	NoClass-3	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:10 M:1.0E-6 S:100
Cluster1	14.00	1.00	0.54	1.00	InC. Classi.:52
Cluster2	28.00	0.32	0.45	2.00	
Cluster3	19.00	0.63	0.80	3.00	
Cluster4	15.00	0.40	0.50	4.00	
Cluster5	1.00	1.00	0.14	NoClass-7	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	19.00	0.32	0.86	7.00	
Cluster8	1.00	1.00	0.08	NoClass-4	
Cluster9	1.00	1.00	0.05	NoClass-2	
Cluster10	1.00	1.00	0.07	NoClass-3	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:100
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:39
Cluster2	24.00	0.46	0.55	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	9.00	1.00	0.75	4.00	
Cluster5	11.00	0.64	0.88	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	23.00	0.26	0.86	7.00	
Cluster8	1.00	1.00	0.14	NoClass-1	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	1.00	1.00	0.05	NoClass-2	
Cluster11	2.00	1.00	0.13	NoClass-3	

Table 10: Best results on ParallelsDS with the optimum N of EM

EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	100.00	0.26	1.00	1.00	InC. Classi.:74

Search for optimum M while N is set to 11

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:2.0E-6 S:100
Cluster1	10.00	1.00	0.38	1.00	InC. Classi.:56
Cluster2	38.00	0.24	0.45	2.00	
Cluster3	7.00	1.00	0.47	3.00	
Cluster4	20.00	0.20	0.33	4.00	
Cluster5	7.00	1.00	0.88	5.00	
Cluster6	11.00	0.45	0.71	6.00	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	2.00	0.50	0.20	8.00	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	2.00	1.00	0.13	NoClass-3	
Cluster11	1.00	1.00	0.05	NoClass-2	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:4.0E-6 S:100
Cluster1	10.00	1.00	0.38	1.00	InC. Classi.:56
Cluster2	38.00	0.24	0.45	2.00	
Cluster3	7.00	1.00	0.47	3.00	
Cluster4	20.00	0.20	0.33	4.00	
Cluster5	7.00	1.00	0.88	5.00	
Cluster6	11.00	0.45	0.71	6.00	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	2.00	0.50	0.20	8.00	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	2.00	1.00	0.13	NoClass-3	
Cluster11	1.00	1.00	0.05	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:8.0E-6 S:100
Cluster1	14.00	1.00	0.54	1.00	InC. Classi.:46
Cluster2	33.00	0.45	0.75	2.00	
Cluster3	8.00	1.00	0.53	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	7.00	1.00	0.88	5.00	
Cluster6	11.00	0.45	0.71	6.00	
Cluster7	20.00	0.15	0.43	7.00	
Cluster8	2.00	0.50	0.20	8.00	
Cluster9	2.00	1.00	0.13	NoClass-3	
Cluster10	1.00	1.00	0.14	NoClass-7	
Cluster11	1.00	1.00	0.05	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:16.0E-6 S:100
Cluster1	11.00	1.00	0.42	1.00	InC. Classi.:48
Cluster2	26.00	0.50	0.65	2.00	
Cluster3	10.00	1.00	0.67	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	8.00	0.88	0.88	5.00	
Cluster6	11.00	0.45	0.71	6.00	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	27.00	0.15	0.80	8.00	
Cluster9	2.00	0.50	0.20	NoClass-8	
Cluster10	2.00	1.00	0.13	NoClass-3	
Cluster11	1.00	1.00	0.05	NoClass-2	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:5.0E-7 S:100
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:40
Cluster2	20.00	0.50	0.50	2.00	
Cluster3	11.00	1.00	0.73	3.00	
Cluster4	9.00	1.00	0.75	4.00	
Cluster5	21.00	0.38	1.00	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	18.00	0.33	0.86	7.00	
Cluster8	1.00	1.00	0.14	NoClass-7	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	1.00	1.00	0.05	NoClass-2	
Cluster11	2.00	1.00	0.13	NoClass-3	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:100
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:39
Cluster2	24.00	0.46	0.55	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	9.00	1.00	0.75	4.00	
Cluster5	11.00	0.64	0.88	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	23.00	0.26	0.86	7.00	
Cluster8	1.00	1.00	0.14	NoClass-1	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	1.00	1.00	0.05	NoClass-2	
Cluster11	2.00	1.00	0.13	NoClass-3	

Table 11: Best results on ParallelsDS with optimum N and M for EM

Search for optimum S while N is set to 11 and M is set to 1.0E-6

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:50
Cluster1	32.00	0.34	0.42	1.00	InC. Classi.:57
Cluster2	32.00	0.31	0.50	2.00	
Cluster3	6.00	1.00	0.40	3.00	
Cluster4	8.00	1.00	0.67	4.00	
Cluster5	5.00	0.80	0.50	5.00	
Cluster6	8.00	0.50	0.57	6.00	
Cluster7	1.00	1.00	0.14	NoClass-6	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.05	NoClass-2	
Cluster10	2.00	1.00	0.13	NoClass-3	
Cluster11	4.00	0.50	0.10	NoClass-2	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:60
Cluster1	11.00	1.00	0.42	1.00	InC. Classi.:47
Cluster2	37.00	0.41	0.75	2.00	
Cluster3	9.00	1.00	0.60	3.00	
Cluster4	10.00	1.00	0.83	4.00	
Cluster5	5.00	0.60	0.38	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	12.00	0.42	0.71	7.00	
Cluster8	1.00	1.00	0.08	NoClass-4	
Cluster9	1.00	1.00	0.07	NoClass-3	
Cluster10	10.00	0.40	0.27	NoClass-3	
Cluster11	3.00	1.00	0.12	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:70
Cluster1	7.00	1.00	0.27	1.00	InC. Classi.:64
Cluster2	21.00	0.38	0.40	2.00	
Cluster3	8.00	1.00	0.53	3.00	
Cluster4	21.00	0.19	0.33	4.00	
Cluster5	20.00	0.20	0.50	5.00	
Cluster6	1.00	1.00	0.14	NoClass-7	
Cluster7	19.00	0.26	0.71	7.00	
Cluster8	1.00	1.00	0.08	NoClass-4	
Cluster9	1.00	1.00	0.07	NoClass-3	
Cluster10	1.00	1.00	0.08	NoClass-4	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:80
Cluster1	9.00	1.00	0.35	1.00	InC. Classi.:59
Cluster2	16.00	0.56	0.45	2.00	
Cluster3	6.00	0.83	0.33	3.00	
Cluster4	5.00	0.80	0.33	4.00	
Cluster5	11.00	0.45	0.63	5.00	
Cluster6	16.00	0.31	0.71	6.00	
Cluster7	3.00	0.67	0.08	NoClass-1	
Cluster8	31.00	0.13	0.80	8.00	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.14	NoClass-6	
Cluster11	1.00	1.00	0.13	NoClass-5	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:90
Cluster1	8.00	1.00	0.31	1.00	InC. Classi.:66
Cluster2	39.00	0.26	0.50	2.00	
Cluster3	15.00	0.53	0.53	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	4.00	0.25	0.13	5.00	
Cluster6	6.00	0.33	0.29	6.00	
Cluster7	7.00	0.29	0.29	7.00	
Cluster8	13.00	0.15	0.40	8.00	
Cluster9	1.00	1.00	0.05	NoClass-2	
Cluster10	5.00	1.00	0.33	NoClass-3	
Cluster11	1.00	1.00	0.13	NoClass-5	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:100
Cluster1	15.00	1.00	0.58	1.00	InC. Classi.:39
Cluster2	24.00	0.46	0.55	2.00	
Cluster3	12.00	1.00	0.80	3.00	
Cluster4	9.00	1.00	0.75	4.00	
Cluster5	11.00	0.64	0.88	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	23.00	0.26	0.86	7.00	
Cluster8	1.00	1.00	0.14	NoClass-1	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	1.00	1.00	0.05	NoClass-2	
Cluster11	2.00	1.00	0.13	NoClass-3	

Table 12: Best results on ParallelsDS with the optimum parameters of EM

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:110
Cluster1	15.00	0.27	0.15	1.00	InC. Classi.:68
Cluster2	55.00	0.25	0.70	2.00	
Cluster3	7.00	0.71	0.33	3.00	
Cluster4	11.00	0.64	0.58	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.07	NoClass-3	
Cluster7	8.00	0.25	0.29	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:120
Cluster1	15.00	0.93	0.54	1.00	InC. Classi.:48
Cluster2	22.00	0.36	0.40	2.00	
Cluster3	3.00	1.00	0.20	3.00	
Cluster4	7.00	0.86	0.50	4.00	
Cluster5	6.00	0.67	0.50	5.00	
Cluster6	9.00	0.67	0.86	6.00	
Cluster7	8.00	0.88	1.00	7.00	
Cluster8	26.00	0.15	0.80	8.00	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	2.00	1.00	0.13	NoClass-3	
Cluster11	1.00	1.00	0.04	NoCLass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:130
Cluster1	2.00	1.00	0.08	1.00	InC. Classi.:69
Cluster2	53.00	0.28	0.75	2.00	
Cluster3	3.00	1.00	0.20	3.00	
Cluster4	8.00	0.50	0.33	4.00	
Cluster5	22.00	0.32	0.88	5.00	
Cluster6	1.00	1.00	0.07	NoClass-3	
Cluster7	6.00	0.17	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	2.00	1.00	0.13	NoClass-3	
Cluster10	1.00	1.00	0.05	NoClass-2	
Cluster11	1.00	1.00	0.08	NoClass-4	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:140
Cluster1	11.00	1.00	0.42	1.00	InC. Classi.:55
Cluster2	32.00	0.28	0.45	2.00	
Cluster3	6.00	1.00	0.40	3.00	
Cluster4	8.00	0.75	0.50	4.00	
Cluster5	1.00	1.00	0.08	NoClass-4	
Cluster6	17.00	0.29	0.71	6.00	
Cluster7	8.00	0.63	0.71	7.00	
Cluster8	14.00	0.21	0.60	8.00	
Cluster9	1.00	1.00	0.05	NoClass-2	
Cluster10	1.00	1.00	0.08	NoClass-4	
Cluster11	1.00	1.00	0.08	NoClass-8	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:150
Cluster1	9.00	1.00	0.35	1.00	InC. Classi.:67
Cluster2	54.00	0.22	0.60	2.00	
Cluster3	4.00	1.00	0.27	3.00	
Cluster4	9.00	0.44	0.33	4.00	
Cluster5	2.00	1.00	0.25	5.00	
Cluster6	1.00	1.00	0.14	NoClass-7	
Cluster7	16.00	0.25	0.57	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.08	NoClass-4	
Cluster10	2.00	1.00	0.07	NoClass-3	
Cluster11	1.00	1.00	0.05	NoClass-2	

Only Answer					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	10.00	0.80	0.31	1.00	InC. Classi.:71
Cluster2	13.00	0.38	0.25	2.00	
Cluster3	75.00	0.20	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.14	6.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:76
Cluster2	93.00	0.22	1.00	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.14	7.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:77
Cluster2	84.00	0.20	0.85	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	9.00	0.22	0.17	4.00	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	9.00	0.56	0.19	1.00	InC. Classi.:76
Cluster2	83.00	0.20	0.85	2.00	
Cluster3	3.00	1.00	0.12	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	8-Jan	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:9 M:1.0E-6 S:100
Cluster1	3.00	1.00	0.12	1.00	InC. Classi.:76
Cluster2	9.00	0.44	0.20	2.00	
Cluster3	82.00	0.18	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:10 M:1.0E-6 S:100
Cluster1	7.00	1.00	0.27	1.00	InC. Classi.:72
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	73.00	0.21	1.00	3.00	
Cluster4	11.00	0.18	0.17	4.00	
Cluster5	1.00	1.00	0.05	NoClass-2	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	2.00	0.50	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.05	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:100
Cluster1	5.00	1.00	0.19	1.00	InC. Classi.:76
Cluster2	3.00	0.67	0.10	2.00	
Cluster3	80.00	0.19	1.00	3.00	
Cluster4	2.00	0.50	0.08	4.00	
Cluster5	2.00	0.50	0.04	NoClass-1	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	2.00	0.50	0.14	7.00	
Cluster8	1.00	1.00	0.05	NoClass-2	
Cluster9	2.00	0.50	0.05	NoClass-2	
Cluster10	1.00	1.00	0.04	NoClass-1	
Cluster11	1.00	1.00	0.04	NoClass-1	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	99.00	0.25	0.96	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.04	NoClass-1	

Question +Answers					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	96.00	0.25	0.92	1.00	InC. Classi.:74
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:6 M:1.0E-6 S:100
Cluster1	93.00	0.23	0.81	1.00	InC. Classi.:77
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	3.00	1.00	0.12	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.14	6.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:7 M:1.0E-6 S:100
Cluster1	92.00	0.24	0.85	1.00	InC. Classi.:75
Cluster2	2.00	1.00	0.10	2.00	
Cluster3	1.00	1.00	0.05	NoClass-2	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	2.00	1.00	0.08	NoClass-1	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:8 M:1.0E-6 S:100
Cluster1	48.00	0.44	0.81	1.00	InC. Classi.:62
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	46.00	0.33	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.04	NoClass-1	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:9 M:1.0E-6 S:100
Cluster1	4.00	1.00	0.15	1.00	InC. Classi.:72
Cluster2	27.00	0.30	0.40	2.00	
Cluster3	64.00	0.23	1.00	3.00	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.13	5.00	
Cluster6	1.00	1.00	0.04	NoClass-1	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.05	NoClass-2	
Cluster9	None				
EM	Size	Precision	Recall	ConsideredAs	I:100 N:10 M:1.0E-6 S:100
Cluster1	4.00	1.00	0.04	1.00	InC. Classi.:77
Cluster2	87.00	0.21	0.90	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	
Cluster5	1.00	1.00	0.05	NoClass-2	
Cluster6	4.00	0.75	0.12	NoClass-1	
Cluster7	1.00	1.00	0.14	7.00	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	None				
Cluster10	None				
EM	Size	Precision	Recall	ConsideredAs	I:100 N:11 M:1.0E-6 S:100
Cluster1	7.00	0.71	0.19	1.00	InC. Classi.:70
Cluster2	73.00	0.25	0.90	2.00	
Cluster3	7.00	0.43	0.20	3.00	
Cluster4	1.00	1.00	0.08	4.00	
Cluster5	6.00	0.33	0.25	5.00	
Cluster6	1.00	1.00	0.14	6.00	
Cluster7	1.00	1.00	0.04	NoClass-1	
Cluster8	1.00	1.00	0.04	NoClass-1	
Cluster9	1.00	1.00	0.04	NoClass-1	
Cluster10	1.00	1.00	0.13	NoClass-5	
Cluster11	1.00	1.00	0.04	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	97.00	0.25	0.92	1.00	InC. Classi.:75
Cluster2	1.00	1.00	0.05	2.00	
Cluster3	1.00	1.00	0.04	NoClass-1	
Cluster4	1.00	1.00	0.04	NoClass-1	

1.8.2-1 GoDaddyDS

Only Question					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	80.00	0.51	1.00	1.00	InC. Classi.: 39
Cluster2	20.00	1.00	0.80	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	29.00	0.90	0.63	1.00	InC. Classi.: 34
Cluster2	48.00	0.54	0.76	2.00	
Cluster3	23.00	0.61	0.56	3.00	

Table 13: Best results on GoDaddyDS with optimum N for EM

EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	20.00	0.65	0.32	1.00	InC. Classi.: 43
Cluster2	63.00	0.54	1.00	2.00	
Cluster3	10.00	1.00	0.40	3.00	
Cluster4	7.00	0.57	0.16	NoClass-3	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	13.00	0.69	0.22	1.00	InC. Classi.: 48
Cluster2	71.00	0.45	0.94	2.00	
Cluster3	12.00	0.92	0.44	3.00	
Cluster4	2.00	1.00	0.05	NoClass-1	
Cluster5	2.00	1.00	0.05	NoClass-1	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	100.00	0.41	1.00	1.00	InC. Classi.: 59

Search for optimum M while N is set to 3

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:5.0E-7 S:100
Cluster1	57.00	0.54	0.76	1.00	InC. Classi.: 40
Cluster2	32.00	0.63	0.59	2.00	
Cluster3	11.00	0.82	0.36	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	29.00	0.90	0.63	1.00	InC. Classi.: 34
Cluster2	48.00	0.54	0.76	2.00	
Cluster3	23.00	0.61	0.56	3.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:100
Cluster1	31.00	0.97	0.73	1.00	InC. Classi.:19
Cluster2	43.00	0.70	0.88	2.00	
Cluster3	26.00	0.81	0.84	3.00	

Table 14: Best results with optimum N and M on GoDaddyDS for EM

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:4.0E-6 S:100
Cluster1	34.00	0.91	0.76	1.00	InC. Classi.: 20
Cluster2	41.00	0.68	0.82	2.00	
Cluster3	25.00	0.84	0.84	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:8.0E-6 S:100
Cluster1	25.00	0.92	0.56	1.00	InC. Classi.:48
Cluster2	13.00	0.46	0.18	2.00	
Cluster3	62.00	0.37	0.92	3.00	

Search for S while N is set to 3 and M is set to 2.0E-6

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:50
Cluster1	36.00	0.69	0.61	1.00	InC. Classi.:30
Cluster2	46.00	0.59	0.79	2.00	
Cluster3	18.00	2.00	0.72	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:60
Cluster1	28.00	0.89	0.61	1.00	InC. Classi.:37
Cluster2	21.00	0.81	0.50	2.00	
Cluster3	51.00	0.41	0.84	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:70
Cluster1	40.00	0.83	0.80	1.00	InC. Classi.:26
Cluster2	45.00	0.64	0.85	2.00	
Cluster3	15.00	0.80	0.48	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:80
Cluster1	2.00	1.00	0.05	1.00	InC. Classi.:47
Cluster2	68.00	0.46	0.91	2.00	
Cluster3	30.00	0.67	0.80	3.00	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:90
Cluster1	21.00	0.67	0.34	1.00	InC. Classi.:54
Cluster2	60.00	0.37	0.65	2.00	
Cluster3	19.00	0.53	0.40	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:100
Cluster1	31.00	0.97	0.73	1.00	InC. Classi.:19
Cluster2	43.00	0.70	0.88	2.00	
Cluster3	26.00	0.81	0.84	3.00	

Table 15: Best results on GoDaddyDS with optimum parameters of EM

EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:110
Cluster1	36.00	0.69	0.61	1.00	InC. Classi.:30
Cluster2	46.00	0.59	0.79	2.00	
Cluster3	18.00	1.00	0.72	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:120
Cluster1	74.00	0.42	0.76	1.00	InC. Classi.:54
Cluster2	20.00	0.50	0.29	2.00	
Cluster3	6.00	1.00	0.24	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:130
Cluster1	42.00	0.60	0.61	1.00	InC. Classi.:52
Cluster2	39.00	0.33	0.38	2.00	
Cluster3	19.00	0.53	0.40	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:140
Cluster1	41.00	0.63	0.63	1.00	InC. Classi.:35
Cluster2	40.00	0.50	0.59	2.00	
Cluster3	19.00	1.00	0.76	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:2.0E-6 S:150
Cluster1	10.00	0.60	0.15	1.00	InC. Classi.:49
Cluster2	45.00	0.51	0.68	2.00	
Cluster3	45.00	0.49	0.88	3.00	

Only Answer					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.: 56
Cluster2	3.00	1.00	0.09	2.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	96.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	95.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Cluster4	1.00	1.00	0.04	NoClass-3	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	93.00	0.43	0.98	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.08	3.00	
Cluster4	1.00	1.00	0.02	NoClass-1	
Cluster5	1.00	1.00	0.03	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	96.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	

Question+Answer					
EM	Size	Precision	Recall	ConsideredAs	I:100 N:2 M:1.0E-6 S:100
Cluster1	97.00	0.42	1.00	1.00	InC. Classi.: 56
Cluster2	3.00	1.00	0.09	2.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:3 M:1.0E-6 S:100
Cluster1	96.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:4 M:1.0E-6 S:100
Cluster1	95.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	
Cluster4	1.00	1.00	0.04	NoClass-3	

EM	Size	Precision	Recall	ConsideredAs	I:100 N:5 M:1.0E-6 S:100
Cluster1	93.00	0.43	0.98	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	2.00	1.00	0.08	3.00	
Cluster4	1.00	1.00	0.02	NoClass-1	
Cluster5	1.00	1.00	0.03	NoClass-2	
EM	Size	Precision	Recall	ConsideredAs	I:100 N:-1 M:1.0E-6 S:100
Cluster1	96.00	0.43	1.00	1.00	InC. Classi.: 55
Cluster2	3.00	1.00	0.09	2.00	
Cluster3	1.00	1.00	0.04	3.00	

1.8.3 Cosine Similarity Results

1.8.3-1 PrincetonDS

PrincetonDS					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	30	0.233	0.35	0.188	1
Cluster2	27	0.333	0.45	0.138	3
Cluster3	5	0.4	0.1	0.075	3
Cluster4	6	0.5	0.15	0.063	4
Cluster5	6	0.333	0.1	0.07	4
Cluster6	5	0.4	0.1	0.059	1
Cluster7	2	0.5	0.05	0.043	1
Cluster8	4	0.75	0.15	0.051	4
Cluster9	2	0.5	0.05	0.037	1
Cluster10	2	0.5	0.05	0.034	1
Cluster11	2	0.5	0.05	0.022	2
Cluster12	2	0.5	0.05	0.03	2
Fwithin:0.0674					
Fbetween:81.7244			Content:0.2 CosineSim.Thresh.:0.05		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	38	0.289	0.55	0.222	5
Cluster2	28	0.393	0.55	0.128	3
Cluster3	3	0.333	0.05	0.059	1
Cluster4	8	0.5	0.2	0.08	4
Cluster5	5	0.4	0.1	0.059	1
Cluster6	2	1	0.1	0.03	2
Cluster7	4	0.5	0.1	0.067	1
Cluster8	3	0.667	0.1	0.034	4
Cluster9	2	0.5	0.05	0.039	1
Fbetween:0.0798					
Fwithin:71.2838			Content:0.3 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	46	0.261	0.6	0.243	5
Cluster2	27	0.37	0.5	0.131	3
Cluster4	11	0.545	0.3	0.099	4
Cluster5	2	0.5	0.05	0.018	1
Cluster6	2	1	0.1	0.03	2
Cluster7	2	0.5	0.05	0.023	4
Cluster8	2	0.5	0.05	0.042	1
Fwithin: 0.0794					
Fbetween: 96.7312			Content:0.4 CosineSim.Thresh.:0.05		

Table 16: Best results on PrincetonDS with optimum content value of Cosine Similarity

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	56	0.268	0.75	0.267	5
Cluster3	2	0.5	0.05	0.058	1
Cluster4	7	0.714	0.25	0.074	4
Cluster5	3	0.333	0.05	0.036	1
Cluster6	2	0.5	0.05	0.023	1
Cluster7	2	0.5	0.05	0.045	1
Fwithin: 0.0866					
Fbetween: 97.4584			Content:0.5 CosineSim.Thresh.:0.05		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	60	0.267	0.8	0.284	5
Cluster2	19	0.368	0.35	0.1	3
Cluster3	2	0.5	0.05	0.034	1
Cluster4	7	0.714	0.25	0.059	4
Cluster5	3	0.333	0.05	0.037	1
Cluster6	3	0.667	0.1	0.027	1
Fwithin: 0.0902					
Fbetween: 76.6158			Content:0.6 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	66	0.242	0.8	0.297	5
Cluster2	13	0.308	0.2	0.094	2
Cluster3	3	0.333	0.05	0.092	1
Cluster4	8	0.5	0.2	0.055	1
Cluster5	3	0.333	0.05	0.057	1
Cluster6	2	0.5	0.05	0.016	4
Cluster7	3	0.667	0.1	0.025	1
Fwithin: 0.0908					
Fbetween: 94.1932			Content:0.7 CosineSim.Thresh.:0.05		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	66	0.242	0.8	0.248	5
Cluster2	16	0.312	0.25	0.102	1
Cluster3	3	0.333	0.05	0.051	1
Cluster4	5	0.8	0.2	0.055	4
Cluster5	2	0.5	0.05	0.026	1
Cluster6	2	0.5	0.05	0.015	4
Fwithin: 0.0863					
Fbetween: 68.1902			Content:0.4 CosineSim.Thresh.:0.03		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	56	0.268	0.75	0.248	5
Cluster2	22	0.364	0.4	0.107	3
Cluster3	2	0.5	0.05	0.056	1
Cluster4	7	0.714	0.25	0.075	4
Cluster5	2	0.5	0.05	0.026	1
Cluster6	2	0.5	0.05	0.015	4
Cluster7	2	0.5	0.05	0.042	1
Fwithin: 0.0814					
Fbetween: 95.0388			Content:0.4 CosineSim.Thresh.:0.04		

Table 17:Best results on PrincetonDS witk the optimum content and Cosine similarity threshold of Cosine Similarity

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	46	0.261	0.6	0.243	5
Cluster2	27	0.37	0.5	0.131	3
Cluster3	2	0.5	0.05	0.05	1
Cluster4	11	0.545	0.3	0.099	4
Cluster5	2	0.5	0.05	0.018	1
Cluster6	2	1	0.1	0.03	2
Cluster7	2	0.5	0.05	0.023	4
Cluster8	2	0.5	0.05	0.042	1
Fwithin: 0.0794					
Fbetween: 96.7312			Content:0.4 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	41	0.293	0.6	0.244	5
Cluster2	27	0.407	0.55	0.126	3
Cluster3	4	0.25	0.05	0.074	1
Cluster4	8	0.5	0.2	0.086	4
Cluster5	3	0.667	0.1	0.039	1
Cluster6	2	1	0.1	0.03	2
Cluster7	3	0.667	0.1	0.048	1
Cluster8	2	0.5	0.05	0.042	1
Cluster9	2	0.5	0.05	0.017	3
Fwithin: 0.0784					
Fbetween:87.3525			Content:0.4 CosineSim.Thresh.:0.06		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	36	0.278	0.5	0.242	5
Cluster2	27	0.407	0.55	0.129	3
Cluster3	3	0.333	0.05	0.06	1
Cluster4	7	0.429	0.15	0.084	4
Cluster5	3	0.667	0.1	0.052	1
Cluster6	2	1	0.1	0.032	2
Cluster7	3	0.333	0.05	0.071	1
Cluster8	6	0.667	0.2	0.058	4
Cluster9	3	0.667	0.1	0.055	4
Cluster10	2	0.5	0.05	0.037	2
Cluster11	2	0.5	0.05	0.017	3
Fwithin: 0.0761					
Fbetween: 94.3987			Content:0.4 CosineSim.Thresh.:0.07		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	43	0.279	0.6	0.243	5
Cluster2	26	0.385	0.5	0.126	3
Cluster3	4	0.25	0.05	0.074	1
Cluster4	10	0.6	0.3	0.094	4
Cluster5	3	0.333	0.05	0.029	1
Cluster6	2	1	0.1	0.03	2
Cluster7	3	0.667	0.1	0.048	1
Cluster8	2	0.5	0.05	0.042	1
Fwithin: 0.0919					
Fbetween:77.8503			Content:0.4 CosineSim.Thresh.:0.055		

1.8.3-2 ParallelsDS

ParallesDS					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	38	0.553	0.808	0.102	1
Cluster2	3	0.667	0.25	0.019	5
Cluster3	5	0.4	0.286	0.038	6
Cluster4	14	0.357	0.25	0.054	2
Cluster5	10	0.4	0.2	0.058	2
Cluster6	2	0.5	0.083	0.031	4
Cluster7	5	0.8	0.333	0.036	4
Cluster8	4	1	0.267	0.017	3
Cluster9	2	1	0.133	0.021	3
Cluster10	3	0.333	0.05	0.026	2
Cluster11	2	0.5	0.143	0.016	6
Cluster12	2	1	0.133	0.006	3
Cluster13	3	0.333	0.067	0.019	3
Fwithin: 0.0342			Content:0.05 CosineSim.Thresh.:0.05		
Fbetween:63.7590					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	39	0.564	0.846	0.114	1
Cluster2	3	0.667	0.25	0.019	5
Cluster3	5	0.4	0.286	0.037	6
Cluster4	14	0.357	0.25	0.054	2
Cluster5	8	0.5	0.2	0.06	2
Cluster6	2	0.5	0.083	0.03	4
Cluster7	5	0.8	0.333	0.035	4
Cluster8	4	1	0.267	0.016	3
Cluster9	2	1	0.133	0.021	3
Cluster10	3	0.333	0.05	0.025	2
Cluster11	2	0.5	0.143	0.016	6
Cluster12	2	1	0.133	0.005	3
Cluster13	3	0.333	0.067	0.018	3
Fwithin:0.0347			Content:0.1 CosineSim.Thresh.:0.05		
Fbetween: 67.1280					

Table 18: Best results on ParallelsDS with the optimum content value for Cosine similarity

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	42	0.571	0.923	0.142	1
Cluster2	3	0.667	0.25	0.017	5
Cluster3	6	0.333	0.077	0.041	1
Cluster4	13	0.308	0.2	0.052	2
Cluster5	6	0.5	0.15	0.052	2
Cluster6	5	0.6	0.429	0.03	6
Cluster7	4	1	0.333	0.03	4
Cluster8	4	1	0.267	0.015	3
Cluster9	3	0.667	0.133	0.019	3
Cluster10	2	1	0.1	0.021	2
Cluster11	2	1	0.133	0.019	3
Fwithin: 0.0398			Content:0.2 CosineSim.Thresh.:0.05		
Fbetween: 63.3966					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	64	0.391	0.962	0.161	1
Cluster2	2	0.5	0.083	0.023	4
Cluster3	3	0.667	0.25	0.023	5
Cluster4	2	0.5	0.125	0.022	5
Cluster5	2	1	0.1	0.027	2
Cluster6	10	0.4	0.2	0.05	2
Cluster7	6	0.5	0.15	0.052	2
Cluster8	3	0.667	0.167	0.024	4
Cluster9	2	1	0.167	0.029	4
Cluster10	2	1	0.133	0.017	3
Fwithin:0.0427					
Fbetween: 92.0497			Content:0.3 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	79	0.329	1	0.183	1
Cluster2	4	0.5	0.1	0.029	2
Cluster3	2	0.5	0.125	0.021	5
Cluster4	4	0.5	0.1	0.031	2
Cluster5	4	0.5	0.1	0.039	2
Fwithin: 0.0607					
Fbetween:64.9993			Content:0.4 CosineSim.Thresh.:0.05		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	82	0.317	1	0.199	1
Cluster2	4	0.5	0.1	0.031	2
Cluster3	2	0.5	0.125	0.02	5
Cluster4	3	1	0.15	0.03	2
Cluster5	2	0.5	0.125	0.019	5
Cluster6	5	0.6	0.15	0.039	2
Fwithin: 0.0564					
Fbetween:94.5244			Content:0.5 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	82	0.317	1	0.213	1
Cluster2	4	0.5	0.1	0.033	2
Cluster3	2	0.5	0.05	0.037	2
Cluster4	4	1	0.2	0.016	2
Cluster5	2	0.5	0.125	0.02	5
Cluster6	4	0.5	0.25	0.038	5
Fwithin: 0.0596					
Fbetween: 103.0724			Content:0.6 CosineSim.Thresh.:0.05		
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	81	0.321	1	0.229	1
Cluster2	4	1	0.2	0.036	2
Cluster3	3	0.667	0.25	0.035	5
Cluster4	4	1	0.2	0.014	2
Cluster5	2	0.5	0.125	0.022	5
Cluster6	3	0.667	0.25	0.02	5
Cluster7	3	0.667	0.25	0.025	5
Fwithin: 0.0544					
Fbetween:123.1413			Content:0.7 CosineSim.Thresh.:0.05		

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	40	0.55	0.846	0.115	1
Cluster2	2	0.5	0.038	0.024	1
Cluster3	3	0.667	0.25	0.019	5
Cluster4	6	0.333	0.077	0.041	1
Cluster5	14	0.357	0.25	0.054	2
Cluster6	6	0.5	0.15	0.053	2
Cluster7	2	0.5	0.083	0.03	4
Cluster8	5	0.8	0.333	0.035	4
Cluster9	4	1	0.267	0.016	3
Cluster10	2	0.5	0.083	0.008	4
Cluster11	2	1	0.1	0.021	2
Cluster12	2	1	0.133	0.021	3
Cluster13	2	0.5	0.143	0.018	6
Cluster14	2	0.5	0.143	0.016	6
Cluster15	2	1	0.133	0.005	3
Fwithin: 0.0319			Content:0.1 CosineSim.Thresh.:0.03		
Fbetween: 80.0028					

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	39	0.564	0.846	0.114	1
Cluster2	3	0.667	0.25	0.019	5
Cluster3	5	0.4	0.286	0.037	6
Cluster4	14	0.357	0.25	0.054	2
Cluster5	8	0.5	0.2	0.06	2
Cluster6	2	0.5	0.083	0.03	4
Cluster7	5	0.8	0.333	0.035	4
Cluster8	4	1	0.267	0.016	3
Cluster9	3	0.667	0.133	0.03	3
Cluster10	2	0.5	0.143	0.018	6
Cluster11	2	0.5	0.143	0.016	6
Cluster12	2	1	0.133	0.005	3
Cluster13	3	0.333	0.067	0.018	3
Fwithin: 0.0349			Content:0.1 CosineSim.Thresh.:0.04		
Fbetween: 66.3920					

Table 19: Best results on ParallelsDS with optimum parameter values for Cosine similarity

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	39	0.564	0.846	0.114	1
Cluster2	3	0.667	0.25	0.019	5
Cluster3	5	0.4	0.286	0.037	6
Cluster4	14	0.357	0.25	0.054	2
Cluster5	8	0.5	0.2	0.06	2
Cluster6	2	0.5	0.083	0.03	4
Cluster7	5	0.8	0.333	0.035	4
Cluster8	4	1	0.267	0.016	3
Cluster9	2	1	0.133	0.021	3
Cluster10	3	0.333	0.05	0.025	2
Cluster11	2	0.5	0.143	0.016	6
Cluster12	2	1	0.133	0.005	3
Cluster13	3	0.333	0.067	0.018	3
Fwithin: 0.0347			Content:0.1 CosineSim.Thresh.:0.045		
Fbetween:67.1280					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	39	0.564	0.846	0.114	1
Cluster2	3	0.667	0.25	0.019	5
Cluster3	5	0.4	0.286	0.037	6
Cluster4	14	0.357	0.25	0.054	2
Cluster5	8	0.5	0.2	0.06	2
Cluster6	2	0.5	0.083	0.03	4
Cluster7	5	0.8	0.333	0.035	4
Cluster8	4	1	0.267	0.016	3
Cluster9	2	1	0.133	0.021	3
Cluster10	3	0.333	0.05	0.025	2
Cluster11	2	0.5	0.143	0.016	6
Cluster12	2	1	0.133	0.005	3
Cluster13	3	0.333	0.067	0.018	3
Fwithin:0.0347			Content:0.1 CosineSim.Thresh.:0.05		
Fbetween: 67.1933					

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	38	0.553	0.808	0.115	
Cluster2	3	0.667	0.25	0.019	
Cluster3	5	0.4	0.286	0.037	
Cluster4	14	0.357	0.25	0.054	
Cluster5	10	0.4	0.2	0.059	
Cluster6	2	0.5	0.083	0.03	
Cluster7	5	0.8	0.333	0.035	
Cluster8	4	1	0.267	0.016	
Cluster9	2	1	0.133	0.021	
Cluster10	3	0.333	0.05	0.025	
Cluster11	2	0.5	0.143	0.016	
Cluster12	2	1	0.133	0.005	
Cluster13	3	0.333	0.067	0.018	
Fwithin:0.0347			Content:0.1 CosineSim.Thresh.:0.06		
Fbetween: 67.1280					

1.8.3-3 GoDaddyDS

GoDaddyDS					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	7	0.857	0.146	0.066	1
Cluster2	33	0.818	0.659	0.097	1
Cluster3	31	0.806	0.735	0.086	2
Cluster4	3	0.333	0.024	0.027	1
Cluster5	2	0.5	0.024	0.018	1
Cluster6	2	0.5	0.029	0.021	2
Cluster7	9	1	0.36	0.033	3
Cluster8	2	1	0.08	0.012	3
Fwithin: 0.0448			Content:0.1 CosineSim.Thresh.:0.05		
Fbetween: 57.9936					

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	9	0.889	0.195	0.08	1
Cluster2	33	0.818	0.659	0.11	1
Cluster3	30	0.833	0.735	0.096	2
Cluster4	2	0.5	0.024	0.019	1
Cluster5	2	0.5	0.029	0.023	2
Cluster6	2	0.5	0.029	0.022	2
Cluster7	9	1	0.36	0.04	3
Cluster8	3	1	0.12	0.019	3
Fwithin: 0.0511			Content:0.2 CosineSim.Thresh.:0.05		
Fbetween:63.4206					

Table 20: Best results on GoDaddyDS with optimum content value for Cosine Similarity

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	16	0.875	0.341	0.101	1
Cluster2	28	0.821	0.561	0.119	1
Cluster3	31	0.806	0.735	0.11	2
Cluster4	3	0.667	0.059	0.029	2
Cluster5	3	1	0.12	0.024	3
Cluster6	6	1	0.24	0.045	3
Cluster7	2	1	0.08	0.014	3
Fwithin: 0.0631			Content:0.3 CosineSim.Thresh.:0.05		
Fbetween:61.3564					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	24	0.917	0.537	0.123	1
Cluster2	24	0.792	0.463	0.126	1
Cluster3	28	0.893	0.735	0.116	2
Cluster4	3	1	0.12	0.04	3
Cluster5	3	0.667	0.059	0.03	2
Cluster6	4	1	16	0.045	3
Fwithin:0.0798			Content:0.4 CosineSim.Thresh.:0.05		
Fbetween:60.5569					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	32	0.75	0.585	0.148	1
Cluster2	21	0.81	0.415	0.132	1
Cluster3	24	0.875	0.618	0.118	2
Cluster4	3	0.667	0.08	0.049	3
Cluster5	6	1	0.24	0.048	3
Cluster6	3	0.667	0.059	0.031	2
Cluster7	2	1	0.08	0.024	3
Fwithin: 0.0786			Content:0.5 CosineSim.Thresh.:0.05		
Fbetween:73.1459					

Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	43	0.674	0.707	0.172	1
Cluster2	17	0.706	0.293	0.13	1
Cluster3	20	0.95	0.559	0.117	2
Cluster4	2	1	0.08	0.025	3
Cluster5	2	0.5	0.029	0.033	2
Cluster6	7	1	0.28	0.057	3
Fwithin: 0.0890			Content:0.6 CosineSim.Thresh.:0.05		
Fbetween: 74.3907					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	46	0.674	0.756	0.19	1
Cluster2	17	0.588	0.244	0.129	1
Cluster3	17	0.941	0.471	0.128	2
Cluster4	9	0.778	0.28	0.079	3
Cluster5	2	0.5	0.029	0.036	2
Cluster6	5	1	0.2	0.043	3
Fwithin: 0.1008			Content:0.7 CosineSim.Thresh.:0.05		
Fbetween: 65.3963					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	21	0.905	0.463	0.103	1
Cluster2	25	0.8	0.488	0.103	1
Cluster3	30	0.833	0.735	0.096	2
Cluster4	3	0.667	0.059	0.028	2
Cluster5	2	1	0.08	0.012	3
Cluster6	5	1	0.2	0.036	3
Cluster7	2	1	0.08	0.012	3
Fwithin: 0.0558			Content:0.2 CosineSim.Thresh.:0.03		
Fbetween: 61.2979					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	13	0.923	0.293	0.087	1
Cluster2	31	0.806	0.61	0.11	1
Cluster3	31	0.806	0.735	0.098	2
Cluster4	3	0.667	0.059	0.03	2
Cluster5	2	1	0.08	0.02	3
Cluster6	8	1	0.32	0.038	3
Cluster7	3	1	0.12	0.019	3
Fwithin:0.0576			Content:0.2 CosineSim.Thresh.:0.04		

Fbetween: 50.9821					
Cosine	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	9	0.889	0.195	0.08	1
Cluster2	33	0.818	0.659	0.11	1
Cluster3	30	0.833	0.735	0.096	2
Cluster4	2	0.5	0.024	0.019	1
Cluster5	2	0.5	0.029	0.023	2
Cluster6	2	0.5	0.029	0.022	2
Cluster7	9	1	0.36	0.04	3
Cluster8	3	1	0.12	0.019	3
Fwithin: 0.0511			Content:0.2 CosineSim.Thresh.:0.05		
Fbetween: 63.4206					

Table 21: Best results on GoDaddyDS with optimum parameters for Cosine Similarity

1.8.4 Jaccard Distance Results

1.8.4-1 PrincetonDS

PrincetonDS					
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	18	0.278	0.25	0.19	5
Cluster2	33	0.303	0.5	0.184	2
Cluster3	4	0.25	0.05	0.061	2
Cluster4	2	0.5	0.05	0.036	2
Cluster5	2	1	0.1	0.036	5
Fwithin: 0.1017					
Fbetween: 65.8845			Content:0.3 JaccardDist.Thresh.:0.994		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	13	0.308	0.2	0.224	4
Cluster2	5	0.4	0.1	0.072	1
Cluster3	24	0.333	0.4	0.186	3
Cluster4	6	0.667	0.2	0.083	2
Cluster5	2	1	0.1	0.041	1
Cluster6	3	0.667	0.1	0.044	2
Cluster7	2	1	0.1	0.048	2
Cluster8	3	0.333	0.05	0.049	3
Cluster9	2	1	0.1	0.046	4
Cluster10	2	0.5	0.05	0.036	4
Cluster11	2	1	0.1	0.036	5
Cluster12	2	1	0.1	0.033	5
Fwithin:0.0748					
Fbetween:123.0724			Content:0.4 JaccardDist.Thresh.:0.994		

Table 22: Best results on PrincetonDS with the optimum content value for Jaccard Distance

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	7	0.429	0.15	0.203	1
Cluster2	6	0.333	0.1	0.149	2
Cluster3	22	0.318	0.35	0.217	5
Cluster4	2	1	0.1	0.039	1
Cluster5	5	0.6	0.15	0.076	2
Cluster6	4	0.75	0.15	0.077	2
Cluster7	2	1	0.1	0.049	4
Cluster8	2	1	0.1	0.028	2
Cluster9	2	1	0.1	0.016	3
Cluster10	3	1	0.15	0.02	3
Cluster11	4	0.5	0.1	0.072	5
Fwithin: 0.0861					
Fbetween: 118.8059			Content:0.5 JaccardDist.Thresh.:0.994		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	3	1	0.15	0.1	1
Cluster2	8	0.375	0.15	0.177	2
Cluster3	15	0.333	0.25	0.25	2
Cluster4	2	0.5	0.05	0.05	4
Cluster5	2	1	0.1	0.032	1
Cluster6	2	1	0.1	0.037	1
Cluster7	2	0.5	0.05	0.061	1
Cluster8	4	0.75	0.15	0.057	2
Cluster9	2	0.5	0.05	0.068	2
Cluster10	4	0.75	0.15	0.09	2
Cluster11	4	1	0.2	0.033	3
Cluster12	4	0.5	0.1	0.081	5
Fwithin: 0.0863					
Fbetween: 143.9731			Content:0.6 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	9	0.333	0.15	0.217	1
Cluster2	9	0.333	0.15	0.272	2
Cluster3	3	0.667	0.1	0.08	4
Cluster4	2	1	0.1	0.032	1
Cluster5	2	1	0.1	0.035	1
Cluster6	3	0.333	0.05	0.078	1
Cluster7	5	1	0.25	0.088	2
Cluster8	2	0.5	0.05	0.071	2
Cluster9	3	1	0.15	0.081	2
Cluster10	3	1	0.15	0.028	3
Cluster11	4	0.5	0.1	0.098	5
Cluster12	2	0.5	0.05	0.024	4
Cluster13	2	1	0.1	0.011	4
Cluster14	2	1	0.1	0.052	5
Fwithin: 0.0833					
Fbetween: 159.8237			Content:0.7 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	13	0.308	0.2	0.224	4
Cluster2	5	0.4	0.1	0.072	1
Cluster3	24	0.333	0.4	0.186	3
Cluster4	6	0.667	0.2	0.083	2
Cluster5	2	1	0.1	0.041	1
Cluster6	3	0.667	0.1	0.044	2
Cluster7	2	1	0.1	0.048	2
Cluster8	3	0.333	0.05	0.049	3
Cluster9	2	1	0.1	0.046	4
Cluster10	2	0.5	0.05	0.036	4
Cluster11	2	1	0.1	0.036	5
Cluster12	2	1	0.1	0.033	5
Fwithin:0.0748					
Fbetween:123.0724			Content:0.4 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	3	1	0.15	0.077	1
Cluster2	5	0.4	0.1	0.135	4
Cluster3	21	0.286	0.3	0.222	2
Cluster4	3	0.333	0.05	0.061	2
Cluster5	3	1	0.15	0.06	2
Cluster6	3	1	0.15	0.027	3
Cluster7	7	0.429	0.15	0.1	4
Fwithin: 0.0974					
Fbetween: 106.3377			Content:0.4 JaccardDist.Thresh.:0.992		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	33	0.242	0.4	0.244	5
Cluster2	7	0.714	0.25	0.088	2
Cluster3	27	0.37	0.5	0.144	3
Cluster4	2	0.5	0.05	0.033	2
Cluster5	2	1	0.1	0.041	1
Cluster6	3	0.667	0.1	0.047	2
Cluster7	2	1	0.1	0.035	4
Fwithin: 0.0902					
Fbetween: 80.2464			Content:0.4 JaccardDist.Thresh.:0.996		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	64	0.25	0.8	0.252	5
Cluster2	10	0.4	0.2	0.078	3
Cluster3	11	0.364	0.2	0.09	1
Cluster4	3	0.667	0.1	0.031	4
Cluster5	2	1	0.1	0.034	1
Fwithin: 0.0971					
Fbetween:49.1687			Content:0.4 JaccardDist.Thresh.:0.998		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	46	0.261	0.6	0.251	5
Cluster2	7	0.429	0.15	0.069	2
Cluster3	20	0.4	0.4	0.122	3
Cluster4	2	1	0.1	0.024	4
Cluster5	2	1	0.1	0.028	1
Cluster6	2	0.5	0.05	0.031	1
Cluster7	4	1	0.2	0.056	1
Cluster8	2	1	0.1	0.027	2
Cluster9	2	0.5	0.05	0.027	4
Cluster10	2	1	0.1	0.019	4
Fwithin: 0.0653					
Fbetween: 108.5668			Content:0.4 JaccardDist.Thresh.:0.997		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	55	0.273	0.75	0.25	5
Cluster2	13	0.385	0.25	0.098	2
Cluster3	8	0.5	0.2	0.087	3
Cluster4	2	1	0.1	0.033	4
Cluster5	4	0.75	0.15	0.039	4
Cluster6	3	1	0.15	0.043	1
Cluster7	2	1	0.1	0.018	1
Cluster8	2	0.5	0.05	0.023	2
Fwithin: 0.0739					
Fbetween: 78.9996		Content:0.4 JaccardDist.Thresh.:0.9975			

Table 23: Best results on PrincetonDS with the optimum parameter values for Jaccard Distance

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	57	0.263	0.75	0.231	5
Cluster2	11	0.364	0.2	0.09	2
Cluster3	9	0.333	0.15	0.087	2
Cluster4	4	1	0.2	0.05	4
Cluster5	2	0.5	0.05	0.024	1
Cluster6	2	1	0.1	0.02	1
Cluster7	4	1	0.2	0.058	1
Fwithin: 0.0800					
Fbetween: 60.6236		Content:0.3 JaccardDist.Thresh.:0.9975			

1.8.4-2 ParallelsDS

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	32	0.625	0.769	0.174	1
Cluster2	3	0.667	0.077	0.036	1
Cluster3	3	0.667	0.1	0.034	2
Cluster4	2	0.5	0.05	0.01	2
Cluster5	4	1	0.267	0.013	3
Cluster6	3	1	0.2	0.007	3
Fwithin : 0.0456					
Fbetween: 117.1710			Content:0.3 JaccardDist.Thresh.:0.994		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	31	0.677	0.808	0.202	1
Cluster2	2	1	0.077	0.023	1
Cluster3	5	0.4	0.1	0.049	2
Cluster4	2	1	0.1	0.017	2
Cluster5	7	1	0.467	0.019	3
Cluster6	4	1	0.267	0.014	3
Cluster7	2	1	0.167	0.011	4
Cluster8	2	1	0.25	0.003	5
Fwithin: 0.0421					
Fbetween:131.3612			Content:0.4 JaccardDist.Thresh.:0.994		

Table 24: Best results on ParallelsDS with optimum content value for Jaccard Distance

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	29	0.69	0.769	0.237	1
Cluster2	6	0.333	0.1	0.057	2
Cluster3	2	0.5	0.038	0.028	1
Cluster4	2	1	0.1	0.015	2
Cluster5	7	1	0.467	0.016	3
Cluster6	4	1	0.267	0.012	3
Cluster7	2	1	0.167	0.01	4
Cluster8	2	1	0.25	0.003	5
Fwithin: 0.0471					
Fbetween: 141.0961			Content:0.5 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	25	0.8	0.769	0.273	1
Cluster2	2	1	0.1	0.036	2
Cluster3	7	0.286	0.1	0.069	2
Cluster4	3	0.667	0.077	0.036	1
Cluster5	2	1	0.1	0.011	2
Cluster6	7	1	0.467	0.014	3
Cluster7	4	1	0.267	0.011	3
Cluster8	2	1	0.167	0.008	4
Cluster9	2	1	0.25	0.002	5
Fwithin: 0.0512					
Fbetween:162.8507			Content:0.6 JaccardDist.Thresh.:0.994		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	21	0.905	0.731	0.298	1
Cluster2	7	0.571	0.2	0.1	2
Cluster3	4	0.75	0.115	0.048	1
Cluster4	2	1	0.1	0.01	2
Cluster5	3	0.667	0.1	0.024	2
Cluster6	7	1	0.467	0.012	3
Cluster7	4	1	0.267	0.009	3
Cluster8	2	1	0.167	0.007	4
Cluster9	2	1	0.25	0.002	5
Cluster10	2	0.5	0.143	0.022	7
Fwithin: 0.0533					
Fbetween:175.7221			Content:0.7 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	21	0.714	0.577	0.219	1
Cluster2	2	0.5	0.038	0.028	1
Cluster3	5	0.4	0.1	0.056	2
Cluster4	2	1	0.077	0.018	1
Cluster5	2	0.5	0.05	0.009	2
Cluster6	2	1	0.133	0.004	3
Cluster7	2	1	0.133	0.002	3
Cluster8	3	1	0.2	0.007	3
Fwithin: 0.0430					
Fbetween: 234.8629			Content:0.4 JaccardDist.Thresh.:0.992		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	31	0.677	0.808	0.202	1
Cluster2	2	1	0.077	0.023	1
Cluster3	5	0.4	0.1	0.049	2
Cluster4	2	1	0.1	0.017	2
Cluster5	7	1	0.467	0.019	3
Cluster6	4	1	0.267	0.014	3
Cluster7	2	1	0.167	0.011	4
Cluster8	2	1	0.25	0.003	5
Fwithin: 0.0421					
Fbetween:131.3612			Content:0.4 JaccardDist.Thresh.:0.994		
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	40	0.625	0.962	0.199	1
Cluster2	5	0.4	0.286	0.039	7
Cluster3	2	1	0.1	0.01	2
Cluster4	2	0.5	0.05	0.009	2
Cluster5	2	1	0.1	0.006	2
Cluster6	2	1	0.1	0.016	2
Cluster7	2	0.5	0.067	0.014	3
Cluster8	10	1	0.667	0.024	3
Cluster9	2	1	0.133	0.005	3
Cluster10	3	1	0.25	0.027	4
Cluster11	2	1	0.167	0.014	4
Cluster12	4	1	0.5	0.022	5
Fwithin: 0.0321					
Fbetween: 138.8941			Content:0.4 JaccardDist.Thresh.:0.996		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	42	0.595	0.962	0.2	1
Cluster2	3	0.667	0.4	0.021	8
Cluster3	3	0.667	0.1	0.024	2
Cluster4	3	0.333	0.05	0.028	2
Cluster5	6	0.667	0.2	0.027	2
Cluster6	2	0.5	0.067	0.014	3
Cluster7	2	1	0.133	0.002	3
Cluster8	8	1	0.533	0.02	3
Cluster9	2	1	0.133	0.005	3
Cluster10	2	1	0.167	0.022	4
Cluster11	2	1	0.167	0.014	4
Cluster12	2	1	0.25	0.003	5
Cluster13	2	1	0.286	0.016	6
Fwithin: 0.0304					
Fbetween: 153.9414			Content:0.4 JaccardDist.Thresh.:0.997		

Table 25: Best results on ParallelsDS with optimum parameter values for Jaccard Distance

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	48	0.521	0.962	0.195	1
Cluster2	4	0.75	0.15	0.036	2
Cluster3	2	0.5	0.125	0.012	5
Cluster4	4	0.5	0.286	0.024	6
Cluster5	9	0.444	0.267	0.037	3
Cluster6	6	0.5	0.15	0.04	2
Cluster7	2	1	0.25	0.003	5
Cluster8	3	0.333	0.083	0.027	4
Cluster9	5	0.8	0.333	0.026	4
Cluster10	4	1	0.267	0.012	3
Cluster11	2	1	0.133	0.016	3
Cluster12	2	0.5	0.05	0.009	2
Cluster13	2	1	0.133	0.004	3
Cluster14	3	0.333	0.067	0.014	3
Fwithin:0.0326					
Fbetween:104.1887			Content:0.4 JaccardDist.Thresh.:0.998		

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	44	0.568	0.962	0.198	1
Cluster2	6	0.5	0.15	0.038	2
Cluster3	3	0.333	0.038	0.03	1
Cluster4	2	0.5	0.143	0.013	7
Cluster5	6	0.667	0.2	0.033	2
Cluster6	2	0.5	0.083	0.022	4
Cluster7	9	0.444	0.267	0.036	3
Cluster8	2	0.5	0.083	0.012	4
Cluster9	2	0.5	0.067	0.016	3
Cluster10	3	0.667	0.1	0.015	2
Cluster11	11	0.818	0.6	0.028	3
Cluster12	2	1	0.25	0.003	5
Fwithin: 0.0371					
Fbetween: 85.7270			Content:0.4 JaccardDist.Thresh.:0.9975		

1.8.4-3 GoDaddyDS

GoDaddyDS					
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	2	1	0.049	0.022	1
Cluster2	11	1	0.268	0.089	1
Cluster3	7	1	0.171	0.059	1
Cluster4	18	0.5	0.265	0.114	2
Cluster5	2	1	0.049	0.03	1
Cluster6	4	0.75	0.073	0.052	1
Cluster7	2	1	0.049	0.011	1
Cluster8	6	0.667	0.16	0.05	3
Cluster9	2	0.5	0.024	0.027	1
Cluster10	3	1	0.088	0.036	2
Cluster11	5	0.6	0.12	0.049	3
Cluster12	6	1	0.176	0.046	2
Cluster13	2	1	0.059	0.01	2
Cluster14	2	1	0.08	0.032	3
Cluster15	5	1	0.2	0.035	3
Cluster16	2	1	0.08	0.012	3
Fwithin:0.0420					
Fbetween:115.6310		Content:0.3 JaccardDist.Thresh.:0.994			

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	2	1	0.049	0.022	1
Cluster2	15	1	0.366	0.109	1
Cluster3	3	1	0.073	0.032	1
Cluster4	18	0.556	0.294	0.13	2
Cluster5	2	1	0.049	0.032	1
Cluster6	2	0.5	0.024	0.032	1
Cluster7	2	1	0.049	0.028	1
Cluster8	2	1	0.049	0.047	1
Cluster9	2	1	0.049	0.014	1
Cluster10	2	0.5	0.029	0.046	2
Cluster11	2	1	0.08	0.025	3
Cluster12	4	1	0.118	0.041	2
Cluster13	3	0.667	0.059	0.03	2
Cluster14	4	0.75	0.12	0.049	3
Cluster15	3	1	0.088	0.024	2
Cluster16	5	1	0.147	0.039	2
Cluster17	3	1	0.088	0.041	2
Cluster18	2	1	0.059	0.021	2
Cluster19	9	1	0.36	0.054	3
Cluster20	2	1	0.08	0.012	
Fwithin:0.0413					
Fbetween:148.0133		Content:0.4 JaccardDist.Thresh.:0.994			

Table 26: Best results on GoDaddyDS with optimum content value for Jaccard Distance

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	3	1	0.073	0.041	1
Cluster2	15	1	0.366	0.121	1
Cluster3	3	0.667	0.049	0.034	1
Cluster4	16	0.5	0.235	0.144	2
Cluster5	3	0.667	0.049	0.044	1
Cluster6	2	1	0.049	0.033	1
Cluster7	2	1	0.049	0.052	1
Cluster8	2	1	0.049	0.017	1
Cluster9	3	0.667	0.059	0.075	2
Cluster10	3	1	0.088	0.037	2
Cluster11	5	0.8	0.118	0.047	2
Cluster12	4	1	0.118	0.033	2
Cluster13	5	1	0.147	0.041	2
Cluster14	2	1	0.059	0.022	2
Cluster15	13	1	0.52	0.07	3
Cluster16	2	1	0.08	0.012	3
Fwithin:0.0515					
Fbetween:142.7847		Content:0.5 JaccardDist.Thresh.:0.994			
Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	3	1	0.073	0.044	1
Cluster2	17	1	0.415	0.122	1
Cluster3	5	1	0.147	0.047	2
Cluster4	7	0.857	0.146	0.124	1
Cluster5	2	1	0.049	0.038	1
Cluster6	4	0.75	0.088	0.081	2
Cluster7	2	1	0.049	0.02	1
Cluster8	3	0.667	0.059	0.085	2
Cluster9	2	0.5	0.024	0.057	1
Cluster10	4	0.75	0.088	0.053	2
Cluster11	2	0.5	0.029	0.033	2
Cluster12	3	0.667	0.049	0.064	1
Cluster13	8	1	0.235	0.054	2
Cluster14	3	1	0.088	0.029	2
Cluster15	2	1	0.059	0.023	2
Cluster16	13	1	0.52	0.072	3
Cluster17	2	1	0.08	0.013	3
Fwithin:0.0564					
Fbetween:173.7739		Content:0.6 JaccardDist.Thresh.:0.994			

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	3	1	0.073	0.047	1
Cluster2	17	1	0.415	0.131	1
Cluster3	6	1	0.176	0.055	2
Cluster4	5	0.8	0.098	0.12	1
Cluster5	2	1	0.049	0.046	1
Cluster6	3	0.667	0.059	0.094	2
Cluster7	8	0.75	0.176	0.124	2
Cluster8	9	1	0.265	0.071	2
Cluster9	3	1	0.088	0.03	2
Cluster10	2	0.5	0.029	0.04	2
Cluster11	2	1	0.059	0.024	2
Cluster12	14	1	0.56	0.082	3
Cluster13	2	1	0.08	0.013	3
Cluster14	2	1	0.08	0.023	3
Fwithin:0.0644					
Fbetween:153.5010		Content:0.7 JaccardDist.Thresh.:0.994			

Jaccard	Size	Precision	Recall	StandardD.	ConsideredAs
Cluster1	2	1	0.049	0.022	1
Cluster2	15	1	0.366	0.109	1
Cluster3	3	1	0.073	0.032	1
Cluster4	18	0.556	0.294	0.13	2
Cluster5	2	1	0.049	0.032	1
Cluster6	2	0.5	0.024	0.032	1
Cluster7	2	1	0.049	0.028	1
Cluster8	2	1	0.049	0.047	1
Cluster9	2	1	0.049	0.014	1
Cluster10	2	0.5	0.029	0.046	2
Cluster11	2	1	0.08	0.025	3
Cluster12	4	1	0.118	0.041	2
Cluster13	3	0.667	0.059	0.03	2
Cluster14	4	0.75	0.12	0.049	3
Cluster15	3	1	0.088	0.024	2
Cluster16	5	1	0.147	0.039	2
Cluster17	3	1	0.088	0.041	2
Cluster18	2	1	0.059	0.021	2
Cluster19	9	1	0.36	0.054	3
Cluster20	2	1	0.08	0.012	
Fwithin:0.0413					
Fbetween:148.0133			Content:0.4 JaccardDist.Thresh.:0.994		

Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	2	1	0.049	0.022	Cons.
Cluster2	6	1	0.146	0.085	1
Cluster3	4	1	0.098	0.047	1
Cluster4	7	0.857	0.146	0.093	1
Cluster5	3	1	0.073	0.049	1
Cluster6	4	0.75	0.088	0.06	1
Cluster7	3	1	0.073	0.037	2
Cluster8	2	0.5	0.029	0.057	1
Cluster9	2	0.5	0.024	0.032	2
Cluster10	2	1	0.049	0.015	1
Cluster11	4	0.5	0.049	0.044	1
Cluster12	3	0.667	0.049	0.051	1
Cluster13	4	1	0.118	0.035	1
Cluster14	5	1	0.147	0.047	2
Cluster15	2	1	0.059	0.021	2
Cluster16	3	1	0.12	0.039	2
Cluster17	11	1	0.44	0.055	3
Cluster18	2	1	0.08	0.012	3
Cluster19	2	1	0.08	0.017	3
Fwithin:0.0430					3
Fbetween:157.4682			Content:0.4 JaccardDist.Thresh.:0.992		

Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	6	1	0.146	0.087	Cons.
Cluster2	22	0.955	0.512	0.119	1
Cluster3	7	0.857	0.176	0.056	1
Cluster4	27	0.63	0.5	0.123	2
Cluster5	3	0.667	0.049	0.041	2
Cluster6	2	0.5	0.024	0.029	1
Cluster7	6	1	0.24	0.05	1
Cluster8	4	0.75	0.12	0.052	3
Cluster9	5	0.8	0.118	0.048	3
Cluster10	2	1	0.08	0.028	2
Cluster11	2	1	0.059	0.011	3
Cluster12	3	1	0.12	0.025	2
Cluster13	2	1	0.08	0.017	3
Fwithin:0.0529					3
Fbetween:90.0508			Content:0.4 JaccardDist.Thresh.:0.996		

Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	9	1	0.22	0.112	Cons.
Cluster2	31	0.871	0.659	0.13	1
Cluster3	19	0.842	0.471	0.11	1
Cluster4	14	0.714	0.294	0.07	2
Cluster5	3	0.667	0.08	0.04	2
Cluster6	4	0.5	0.049	0.045	3
Cluster7	4	0.75	0.12	0.046	1
Cluster8	4	0.75	0.12	0.029	3
Cluster9	2	1	0.08	0.029	3
Fwithin:0.0681					3
Fbetween:64.1584			Content:0.4 JaccardDist.Thresh.:0.997		
Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	24	0.75	0.439	0.14	Cons.
Cluster2	31	0.677	0.512	0.129	1
Cluster3	22	0.909	0.588	0.104	1
Cluster4	8	0.875	0.28	0.054	2
Cluster5	3	0.333	0.024	0.028	3
Cluster6	2	0.5	0.024	0.029	1
Fwithin:0.0807					1
Fbetween:53.3917			Content:0.4 JaccardDist.Thresh.:0.998		
Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	57	0.596	0.829	0.17	Cons.
Cluster2	26	0.385	0.4	0.111	1
Cluster3	8	0.75	0.176	0.054	3
Cluster4	4	1	0.16	0.031	2
Cluster5	4	0.75	0.088	0.034	3
Fwithin:0.0799					2
Fbetween:42.6114			Content:0.4 JaccardDist.Thresh.:0.999		

Jaccard	Size	Precision	Recall	StandartD.	
Cluster1	12	0.833	0.244	0.125	Cons.
Cluster2	34	0.824	0.683	0.131	1
Cluster3	23	0.87	0.588	0.109	1
Cluster4	10	0.6	0.24	0.064	2
Cluster5	2	0.5	0.029	0.028	3
Cluster6	2	1	0.049	0.019	2
Cluster7	5	0.8	0.16	0.043	1
Cluster8	3	0.667	0.08	0.027	3
Cluster9	2	0.5	0.029	0.024	3
Fwithin:0.0635					2
Fbetween:79.9290			Content:0.4 JaccardDist.Thresh.:0.9975		

Table 27: Best results on GoDaddyDS with optimum parameter values for Jaccard Distance

1.8.5 CES Results

1.8.5-1 PrincetonDS

PrincetonDS						
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.024	3	
TOTAL	Elapsed	Time	for	CES	:	91549 ms
Fwithin	:	0.02403995				
Fbetween	:	NaN		DefaultValues		
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.026	3	
TOTAL	Elapsed	Time	for	CES	:	88998 ms
Fwithin	:	0.02565594				
Fbetween	:	NaN		DefaultValues+content=0.4		
DefaultValues+content=0.6 Deadlock						
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.027	3	
TOTAL	Elapsed	Time	for	CES	:	98625 ms
Fwithin	:	0.02727193				
Fbetween	:	NaN	DefaultValues+content=0.3			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.026	3	
TOTAL	Elapsed	Time	for	CES	:	87036 ms
Fwithin	:	0.02565594				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.005			
CES	Size	Precision	Recall	StandartD.	Considered	
TOTAL	Elapsed	Time	for	CES	:	88714 ms
Fwithin	:	NaN				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.007			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.072	5	
-->CLUSTER.2	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES	:	88073 ms
Fwithin	:	0.05386793				
Fbetween	:	182.817779	DefaultValues+content=0.4+tfidf=0.004			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	4	0.5	0.1	0.078	5	
-->CLUSTER.2	4	0.5	0.1	0.049	2	
TOTAL	Elapsed	Time	for	CES	:	87304 ms
Fwithin	:	0.06332931				
Fbetween	:	123.186306	DefaultValues+content=0.4+tfidf=0.003			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES	:	88012 ms
Fwithin	:	0.03596895				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.0045			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.025	5	
-->CLUSTER.2	2	0.5	0.05	0.056	4	
TOTAL	Elapsed	Time	for	CES	:	39024 ms
Fwithin	:	0.0405017				
Fbetween	:	201.614314	DefaultValues+content=0.4+tfidf=0.004+CC=0.2			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.05	5	
-->CLUSTER.2	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES	:	151839 ms
Fwithin	:	0.04312024				
Fbetween	:	142.125542	DefaultValues+content=0.4+tfidf=0.004+CC=0.4			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.026	3	
-->CLUSTER.2	3	0.333	0.05	0.075	3	
TOTAL	Elapsed	Time	for	CES	:	9296 ms
Fwithin	:	0.0501239				
Fbetween	:	130.820572	DefaultValues+content=0.4+tfidf=0.004+CC=0.1			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	4	0.25	0.05	0.071	1	
TOTAL	Elapsed	Time	for	CES	:	46207 ms
Fwithin	:	0.07140343				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.002			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	0.5	0.05	0.045	4	
-->CLUSTER.2	2	0.5	0.05	0.056	4	
-->CLUSTER.3	2	1	0.1	0.027	3	
TOTAL	Elapsed	Time	for	CES	:	35259 ms
Fwithin	:	0.04269835				
Fbetween	:	183.577571	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.004			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	0.409	0.45	0.159	5	
-->CLUSTER.2	19	0.421	0.4	0.15	1	
-->CLUSTER.3	10	0.4	0.2	0.119	4	
-->CLUSTER.4	5	0.4	0.1	0.062	1	
-->CLUSTER.5	10	0.8	0.4	0.095	2	
TOTAL	Elapsed	Time	for	CES	:	46488 ms
Fwithin	:	0.11690936				
Fbetween	:	42.7394327	content=0.4+tfidf=0.004+CC=0.2+E=0.002+S=0.001			

Table 28: Best results on PrincetonDS with optimum parameter values for CES

1.8.5-2 ParallelsDS

ParallelsDS						
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.044	1	
-->CLUSTER.2	12	0.583	0.35	0.051	2	
-->CLUSTER.3	3	0.667	0.1	0.039	2	
-->CLUSTER.4	2	1	0.133	0.001	3	
TOTAL	Elapsed	Time	for	CES	:	14540 ms
Fwithin	:	0.03385617				
Fbetween	:	159.080725	DefaultValues			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.035	1	
-->CLUSTER.2	7	0.571	0.267	0.029	3	
-->CLUSTER.3	10	0.4	0.2	0.039	2	
-->CLUSTER.4	7	0.714	0.25	0.06	2	
TOTAL	Elapsed	Time	for	CES	:	18018 ms
Fwithin	:	0.04096013				
Fbetween	:	117.216838	DefaultValues+content=0.4			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.053	1	
-->CLUSTER.2	10	0.5	0.333	0.022	3	
-->CLUSTER.3	8	0.875	0.35	0.037	2	
-->CLUSTER.4	2	1	0.286	0.031	7	
TOTAL	Elapsed	Time	for	CES	:	14139 ms
Fwithin	:	0.0355969				
Fbetween	:	180.557204	DefaultValues+content=0.6			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.061	1	
-->CLUSTER.2	10	0.7	0.467	0.014	3	
-->CLUSTER.3	5	0.6	0.15	0.014	2	
TOTAL	Elapsed	Time	for	CES	:	10464 ms
Fwithin	:	0.02960522				
Fbetween	:	194.914829	DefaultValues+content=0.7			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	5	0.4	0.077	0.053	1	
-->CLUSTER.2	12	0.75	0.6	0.041	3	
-->CLUSTER.3	11	0.727	0.4	0.044	2	
-->CLUSTER.4	9	0.778	0.583	0.05	4	
TOTAL	Elapsed	Time	for	CES	:	26677 ms
Fwithin	:	0.04702829				
Fbetween	:	45.9026026	DefaultValues+content=0.3			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.4	0.012	8	
-->CLUSTER.2	10	0.7	0.35	0.047	2	
-->CLUSTER.3	3	0.667	0.286	0.032	6	
-->CLUSTER.4	5	1	0.333	0.031	3	
-->CLUSTER.5	6	1	0.5	0.044	4	
TOTAL	Elapsed	Time	for	CES	:	38111 ms
Fwithin	:	0.03321605				
Fbetween	:	66.4168788	DefaultValues+content=0.2			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	11	0.545	0.3	0.055	2	
-->CLUSTER.2	4	1	0.267	0.031	3	
TOTAL	Elapsed	Time	for	CES	:	38016 ms
Fwithin	:	0.04325622				
Fbetween	:	37.1555969	DefaultValues+content=0.1			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	9	0.444	0.154	0.065	1	
-->CLUSTER.2	12	0.75	0.6	0.041	3	
-->CLUSTER.3	12	0.75	0.45	0.048	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES	:	26214 ms
Fwithin	:	0.05120367				
Fbetween	:	37.6689734	DefaultValues+content=0.3+tfidf=0.005			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	11	0.727	0.4	0.044	2	
-->CLUSTER.2	5	0.8	0.333	0.04	4	
TOTAL	Elapsed	Time	for	CES	:	25795 ms
Fwithin	:	0.04232717				
Fbetween	:	38.9908505	DefaultValues+content=0.3+tfidf=0.007			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	12	0.417	0.192	0.07	1	
-->CLUSTER.2	15	0.733	0.733	0.044	3	
-->CLUSTER.3	13	0.769	0.5	0.054	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES	:	25985 ms
Fwithin	:	0.05456437				
Fbetween	:	34.039564	DefaultValues+content=0.3+tfidf=0.004			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.731	0.163	1	
-->CLUSTER.2	16	0.688	0.733	0.045	3	
-->CLUSTER.3	15	0.8	0.6	0.059	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES	:	26037 ms
Fwithin	:	0.07940455				
Fbetween	:	25.8438552	DefaultValues+content=0.3+tfidf=0.003			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	29	0.655	0.731	0.148	1	
-->CLUSTER.2	19	0.632	0.6	0.067	2	
-->CLUSTER.3	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES	:	12383 ms
Fwithin	:	0.08879105				
Fbetween	:	25.0069267	DefaultValues+content=0.3+tfidf=0.003+CC=0.2			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	36	0.583	0.808	0.172	1	
-->CLUSTER.2	15	0.733	0.55	0.049	2	
-->CLUSTER.3	7	0.714	0.417	0.059	4	
-->CLUSTER.4	4	1	0.267	0.025	3	
TOTAL	Elapsed	Time	for	CES	:	47421 ms
Fwithin	:	0.07619204				
Fbetween	:	30.7665364	DefaultValues+content=0.3+tfidf=0.003+CC=0.4			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	13	0.385	0.192	0.068	1	
-->CLUSTER.2	16	0.5	0.4	0.06	2	
-->CLUSTER.3	5	1	0.333	0.028	3	
TOTAL	Elapsed	Time	for	CES	:	132933 ms
Fwithin	:	0.05191907				
Fbetween	:	38.3944264	DefaultValues+content=0.3+tfidf=0.003+CC=0.6			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	14	0.357	0.192	0.071	1	
-->CLUSTER.2	13	0.538	0.35	0.053	2	
-->CLUSTER.3	5	1	0.333	0.028	3	
-->CLUSTER.4	6	1	0.5	0.039	4	
TOTAL	Elapsed	Time	for	CES	:	141874 ms
Fwithin	:	0.04787114				
Fbetween	:	40.040978	DefaultValues+content=0.3+tfidf=0.003+CC=0.7			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	30	0.533	0.615	0.174	1	
-->CLUSTER.2	10	0.4	0.2	0.052	2	
-->CLUSTER.3	6	0.333	0.077	0.045	1	
-->CLUSTER.4	24	0.417	0.667	0.067	3	
-->CLUSTER.5	5	0.4	0.286	0.037	7	
TOTAL	Elapsed	Time	for	CES	:	3136 ms
Fwithin	:	0.07484956				
Fbetween	:	29.3105081	DefaultValues+content=0.3+tfidf=0.003+CC=0.1			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	21	0.381	0.308	0.091	1	
-->CLUSTER.2	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES	:	22943 ms
Fwithin	:	0.07141736				
Fbetween	:	27.2926539	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.02			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	25	0.32	0.308	0.089	1	
-->CLUSTER.2	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES	:	29504 ms
Fwithin	:	0.07024846				
Fbetween	:	24.4973725	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.001			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	26	0.769	0.769	0.154	1	
-->CLUSTER.2	14	0.786	0.55	0.064	2	
-->CLUSTER.3	2	0.5	0.05	0.018	2	
-->CLUSTER.4	24	0.542	0.867	0.053	3	
TOTAL	Elapsed	Time	for	CES	:	10408 ms
Fwithin	:	0.07219857				
Fbetween	:	32.5883863	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.004			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	0.792	0.731	0.152	1	
-->CLUSTER.2	14	0.786	0.55	0.064	2	
-->CLUSTER.3	2	0.5	0.05	0.018	2	
-->CLUSTER.4	24	0.542	0.867	0.053	3	
TOTAL	Elapsed	Time	for	CES	:	9576 ms
Fwithin	:	0.07172974				
Fbetween	:	32.6839483	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.005			
DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.002 - Deadlock						
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	0.467	0.269	0.105	1	
-->CLUSTER.2	3	1	0.2	0.023	3	
TOTAL	Elapsed	Time	for	CES	:	98222 ms
Fwithin	:	0.06381729				
Fbetween	:	45.2150538	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.001			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	28	0.679	0.731	0.151	1	
-->CLUSTER.2	19	0.737	0.7	0.061	2	
-->CLUSTER.3	13	0.769	0.833	0.067	4	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	6	0.833	0.333	0.032	3	
TOTAL	Elapsed	Time	for	CES	:	42038 ms
Fwithin	:	0.06959788				
Fbetween	:	31.7307039	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.004			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	1	0.577	0.152	1	
-->CLUSTER.2	24	0.625	0.75	0.067	2	
-->CLUSTER.3	13	0.923	1	0.071	4	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	4	1	0.267	0.025	3	
TOTAL	Elapsed	Time	for	CES	:	39925 ms
Fwithin	:	0.07028005				
Fbetween	:	36.5533538	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.005			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	1	0.577	0.152	1	
-->CLUSTER.2	16	0.625	0.5	0.06	2	
-->CLUSTER.3	3	0.333	0.05	0.032	2	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	4	1	0.267	0.025	3	
-->CLUSTER.6	10	1	0.833	0.056	4	
TOTAL	Elapsed	Time	for	CES	:	36990 ms
Fwithin	:	0.06020539				
Fbetween	:	44.7976622	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.006			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	6	0.833	0.714	0.036	6	
-->CLUSTER.6	13	1	0.867	0.034	3	
-->CLUSTER.7	5	1	0.625	0.028	5	
TOTAL	Elapsed	Time	for	CES	:	39238 ms
Fwithin	:	0.06378233				
Fbetween	:	39.9852981	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0001			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	1	0.115	0.163	1	
-->CLUSTER.2	9	0.556	0.25	0.041	2	
TOTAL	Elapsed	Time	for	CES	:	39222 ms
Fwithin	:	0.10208461				
Fbetween	:	77.4258518	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.003			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	14	1	0.933	0.033	3	
-->CLUSTER.7	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES	:	39831 ms
Fwithin	:	0.06327724				
Fbetween	:	39.9750019	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0009			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	14	1	0.933	0.033	3	
-->CLUSTER.7	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES	:	39573 ms
Fwithin	:	0.06327724				
Fbetween	:	39.9750019	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0008			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	1	0.923	0.168	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	2	0.5	0.05	0.017	2	
-->CLUSTER.7	14	1	0.933	0.033	3	
-->CLUSTER.8	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES	:	39228 ms
Fwithin	:	0.0590203				
Fbetween	:	46.6146868	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0005			

Table 29: Best results on ParallelsDS woth optimum parameter values for CES

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	1	0.923	0.168	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	2	0.5	0.05	0.017	2	
-->CLUSTER.7	14	1	0.933	0.033	3	
-->CLUSTER.8	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES	:	39547 ms
Fwithin	:	0.0590203				
Fbetween	:	46.6146868	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0001			

1.8.5-3 GoDaddyDS

GoDaddyDS						
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	10	0.9	0.22	0.129	1	
-->CLUSTER.2	34	0.588	0.488	0.157	1	
-->CLUSTER.3	5	0.8	0.16	0.065	3	
-->CLUSTER.4	9	1	0.36	0.064	3	
TOTAL	Elapsed	Time	for	CES	:	17372 ms
Fwithin	:	0.10367349				
Fbetween	:	47.237956	DefaultValues.txt			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.463	0.146	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	13	1	0.52	0.087	3	
TOTAL	Elapsed	Time	for	CES	:	27831 ms
Fwithin	:	0.09222384				
Fbetween	:	48.7220668	DefaultValues+content=0.4			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	31	0.548	0.415	0.182	1	
-->CLUSTER.2	15	1	0.6	0.096	3	
-->CLUSTER.3	3	0.333	0.024	0.051	1	
TOTAL	Elapsed	Time	for	CES	:	22150 ms
Fwithin	:	0.10999654				
Fbetween	:	55.4317021	DefaultValues+content=0.6			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.463	0.128	1	
-->CLUSTER.2	3	0.333	0.024	0.04	1	
-->CLUSTER.3	13	1	0.52	0.082	3	
TOTAL	Elapsed	Time	for	CES	:	30314 ms
Fwithin	:	0.08330776				
Fbetween	:	44.422459	DefaultValues+content=0.3			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	15	1	0.6	0.088	3	
TOTAL	Elapsed	Time	for	CES	:	26898 ms
Fwithin	:	0.09731533				
Fbetween	:	41.6826159	DefaultValues+content=0.4+tfidf=0.005			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	31	0.548	0.415	0.146	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	13	1	0.52	0.087	3	
TOTAL	Elapsed	Time	for	CES	:	26887 ms
Fwithin	:	0.09208441				
Fbetween	:	49.5678461	DefaultValues+content=0.4+tfidf=0.007			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	16	1	0.64	0.092	3	
TOTAL	Elapsed	Time	for	CES	:	27834 ms
Fwithin	:	0.09853628				
Fbetween	:	41.2808373	DefaultValues+content=0.4+tfidf=0.004			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.596	0.756	0.163	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	16	1	0.64	0.092	3	
TOTAL	Elapsed	Time	for	CES	:	27171 ms
Fwithin	:	0.09945965				
Fbetween	:	41.1223687	DefaultValues+content=0.4+tfidf=0.003			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	15	1	0.6	0.088	3	
TOTAL	Elapsed	Time	for	CES	:	27714 ms
Fwithin	:	0.09731533				
Fbetween	:	41.6826159	DefaultValues+content=0.4+tfidf=0.0045			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	46	0.63	0.707	0.159	1	
-->CLUSTER.2	16	0.938	0.6	0.09	3	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
TOTAL	Elapsed	Time	for	CES	:	13328 ms
Fwithin	:	0.0943803				
Fbetween	:	47.9574899	DefaultValues+content=0.4+tfidf=0.004+CC=0.2			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	42	0.714	0.882	0.154	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	12	1	0.48	0.081	3	
TOTAL	Elapsed	Time	for	CES	:	35558 ms
Fwithin	:	0.11595811				
Fbetween	:	37.9654753	DefaultValues+content=0.4+tfidf=0.004+CC=0.4			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	32	0.938	0.732	0.14	1	
-->CLUSTER.2	11	0.636	0.28	0.082	3	
-->CLUSTER.3	29	0.862	0.735	0.107	2	
-->CLUSTER.4	9	1	0.36	0.067	3	
TOTAL	Elapsed	Time	for	CES	:	52678 ms
Fwithin	:	0.09906089				
Fbetween	:	37.5409833	DefaultValues+content=0.4+tfidf=0.004+CC=0.5			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.842	0.78	0.151	1	
-->CLUSTER.2	23	0.826	0.76	0.102	3	
-->CLUSTER.3	5	0.6	0.073	0.059	1	
-->CLUSTER.4	16	1	0.471	0.076	2	
TOTAL	Elapsed	Time	for	CES	:	84151 ms
Fwithin	:	0.09712504				
Fbetween	:	40.0123355	DefaultValues+content=0.4+tfidf=0.004+CC=0.6			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	32	0.719	0.561	0.138	1	
-->CLUSTER.2	17	0.765	0.52	0.096	3	
-->CLUSTER.3	2	0.5	0.024	0.018	1	
-->CLUSTER.4	16	1	0.471	0.078	2	
TOTAL	Elapsed	Time	for	CES	:	159545 ms
Fwithin	:	0.08248866				
Fbetween	:	43.1753983	DefaultValues+content=0.4+tfidf=0.004+CC=0.7			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	40	0.725	0.707	0.149	1	
-->CLUSTER.2	16	0.625	0.294	0.096	2	
-->CLUSTER.3	5	1	0.147	0.043	2	
TOTAL	Elapsed	Time	for	CES	:	196469 ms
Fwithin	:	0.09595713				
Fbetween	:	25.6707884	DefaultValues+content=0.4+tfidf=0.004+CC=0.8			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	54	0.593	0.78	0.167	1	
-->CLUSTER.2	4	0.5	0.049	0.053	1	
-->CLUSTER.3	5	1	0.2	0.04	3	
TOTAL	Elapsed	Time	for	CES	:	10807 ms
Fwithin	:	0.08639169				
Fbetween	:	47.5998202	DefaultValues+content=0.4+tfidf=0.004+CC=0.1			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.596	0.756	0.167	1	
-->CLUSTER.2	15	0.933	0.56	0.09	3	
TOTAL	Elapsed	Time	for	CES	:	20075 ms
Fwithin	:	0.12844943				
Fbetween	:	33.4529307	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.002			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	42	0.619	0.634	0.156	1	
-->CLUSTER.2	8	0.5	0.098	0.073	1	
-->CLUSTER.3	16	0.938	0.6	0.09	3	
-->CLUSTER.4	2	0.5	0.024	0.034	1	
TOTAL	Elapsed	Time	for	CES	:	12199 ms
Fwithin	:	0.08836014				
Fbetween	:	49.0565863	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.004			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.526	0.488	0.15	1	
-->CLUSTER.2	15	0.933	0.56	0.09	3	
TOTAL	Elapsed	Time	for	CES	:	31969 ms
Fwithin	:	0.11992432				
Fbetween	:	39.070154	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.001			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	28	0.536	0.441	0.146	2	
-->CLUSTER.2	25	0.92	0.561	0.115	1	
-->CLUSTER.3	13	1	0.52	0.085	3	
TOTAL	Elapsed	Time	for	CES	:	37590 ms
Fwithin	:	0.1153912				
Fbetween	:	40.4554957	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.002			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	37	0.703	0.765	0.149	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
-->CLUSTER.4	13	1	0.52	0.085	3	
TOTAL	Elapsed	Time	for	CES	:	32803 ms
Fwithin	:	0.09523945				
Fbetween	:	49.0641692	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.004			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.538	0.683	0.165	1	
-->CLUSTER.2	10	0.7	0.171	0.093	1	
-->CLUSTER.3	14	1	0.56	0.08	3	
TOTAL	Elapsed	Time	for	CES	:	51257 ms
Fwithin	:	0.11271651				
Fbetween	:	37.6016151	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.001			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.684	0.765	0.152	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
-->CLUSTER.4	12	1	0.48	0.081	3	
TOTAL	Elapsed	Time	for	CES	:	33341 ms
Fwithin	:	0.0951172				
Fbetween	:	48.8829746	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.0035			

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
TOTAL	Elapsed	Time	for	CES	:	35297 ms
Fwithin	:	0.11810412				
Fbetween	:	37.5378643	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.001			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	36	0.694	0.735	0.139	2	
-->CLUSTER.2	21	0.905	0.463	0.108	1	
-->CLUSTER.3	6	1	0.24	0.07	3	
TOTAL	Elapsed	Time	for	CES	:	35156 ms
Fwithin	:	0.10533146				
Fbetween	:	36.0946898	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.003			
CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
-->CLUSTER.4	2	1	0.08	0.012	3	
TOTAL	Elapsed	Time	for	CES	:	35336 ms
Fwithin	:	0.09167867				
Fbetween	:	45.957703	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.0009			

Table 30: Best results on GoDaddyDS with the optimum parameter values for CES

CES	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
-->CLUSTER.4	2	1	0.08	0.012	3	
TOTAL Elapsed Time for CES : 35337 ms						
Fwithin : 0.0916786674710695			content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.0008			
Fbetween : 45.9577029848812						

1.8.6 CES+ Results

1.8.6-1 PrincetonDS

PrincetonDS						
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.024	3	
TOTAL	Elapsed	Time	for	CES+	:	1876 ms
Fwithin	:	0.02403995				
Fbetween	:	NaN		DefaultValues		
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.026	3	
TOTAL	Elapsed	Time	for	CES+	:	1827 ms
Fwithin	:	0.02565594				
Fbetween	:	NaN		DefaultValues+content=0.4		
CES+	Size	Precision	Recall	StandartD.	Considered	
TOTAL	Elapsed	Time	for	CES+	:	1739 ms
Fwithin	:	NaN				
Fbetween	:	NaN	DefaultValues+content=0.6			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.027	3	
TOTAL	Elapsed	Time	for	CES+	:	1911 ms
Fwithin	:	0.02727193				
Fbetween	:	NaN	DefaultValues+content=0.3			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.026	3	

TOTAL	Elapsed	Time	for	CES+	:	1803 ms
Fwithin	:	0.02565594				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.005			
CES+	Size	Precision	Recall	StandartD.	Considered	
TOTAL	Elapsed	Time	for	CES+	:	1795 ms
Fwithin	:	NaN				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.007			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.072	5	
-->CLUSTER.2	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES+	:	1820 ms
Fwithin	:	0.05386793				
Fbetween	:	182.817779	DefaultValues+content=0.4+tfidf=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	4	0.5	0.1	0.078	5	
-->CLUSTER.2	4	0.5	0.1	0.049	2	
TOTAL	Elapsed	Time	for	CES+	:	1829 ms
Fwithin	:	0.06332931				
Fbetween	:	123.186306	DefaultValues+content=0.4+tfidf=0.003			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES+	:	1800 ms
Fwithin	:	0.03596895				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.0045			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.1	0.025	5	
-->CLUSTER.2	2	0.5	0.05	0.056	4	
TOTAL	Elapsed	Time	for	CES+	:	1073 ms
Fwithin	:	0.0405017				
Fbetween	:	201.614314	DefaultValues+content=0.4+tfidf=0.004+CC=0.2			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	0.667	0.1	0.05	5	
-->CLUSTER.2	3	0.667	0.1	0.036	3	
TOTAL	Elapsed	Time	for	CES+	:	2561 ms
Fwithin	:	0.04312024				
Fbetween	:	142.125542	DefaultValues+content=0.4+tfidf=0.004+CC=0.4			

CES+	Size	Precision	Recall	StandartD.	Considered	

-->CLUSTER.1	2	1	0.1	0.026	3	
-->CLUSTER.2	3	0.333	0.05	0.075	3	
TOTAL	Elapsed	Time	for	CES+	:	498 ms
Fwithin	:	0.0501239				
Fbetween	:	130.820572	DefaultValues+content=0.4+tfidf=0.004+CC=0.1			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	0.5	0.05	0.039	1	
TOTAL	Elapsed	Time	for	CES+	:	2354 ms
Fwithin	:	0.0393				
Fbetween	:	NaN				
			DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.001			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	4	0.25	0.05	0.071	1	
TOTAL	Elapsed	Time	for	CES+	:	1230 ms
Fwithin	:	0.07140343				
Fbetween	:	NaN	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.002			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	0.5	0.05	0.045	4	
-->CLUSTER.2	2	0.5	0.05	0.056	4	
-->CLUSTER.3	2	1	0.1	0.027	3	
TOTAL	Elapsed	Time	for	CES+	:	1050 ms
Fwithin	:	0.04269835				
Fbetween	:	183.577571	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.004			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	0.409	0.45	0.159	5	
-->CLUSTER.2	19	0.421	0.4	0.15	1	
-->CLUSTER.3	10	0.4	0.2	0.119	4	
-->CLUSTER.4	5	0.4	0.1	0.062	1	
-->CLUSTER.5	10	0.8	0.4	0.095	2	
TOTAL	Elapsed	Time	for	CES+	:	1309 ms
Fwithin	:	0.11690936				
Fbetween	:	42.7394327	content=0.4+tfidf=0.004+CC=0.2+E=0.002+S=0.001			

Table 31: Best results on PrincetonDS with the optimum parameter values for CES+

CES+	Size	Precision	Recall	StandartD.	Considered	
TOTAL	Elapsed	Time	for	CES+	:	1242 ms
Fwithin	:	NaN				
Fbetween	:	NaN	content=0.4+tfidf=0.004+CC=0.2+E=0.002+S=0.003			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	25	0.36	0.45	0.158	5	
-->CLUSTER.2	19	0.421	0.4	0.15	1	
-->CLUSTER.3	11	0.364	0.2	0.118	4	
-->CLUSTER.4	6	0.5	0.15	0.102	5	
-->CLUSTER.5	10	0.8	0.4	0.095	2	
TOTAL	Elapsed	Time	for	CES+	:	1316 ms
Fwithin	:	0.12475889				
Fbetween	:	42.3017076	content=0.4+tfidf=0.004+CC=0.2+E=0.002+S=0.0009			

1.8.6-2 ParallelsDS

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.044	1	
-->CLUSTER.2	12	0.583	0.35	0.051	2	
-->CLUSTER.3	3	0.667	0.1	0.039	2	
-->CLUSTER.4	2	1	0.133	0.001	3	
TOTAL	Elapsed	Time	for	CES+	:	444 ms
Fwithin	:	0.03385617				
Fbetween	:	159.080725	DefaultValues			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.035	1	
-->CLUSTER.2	7	0.571	0.267	0.029	3	
-->CLUSTER.3	10	0.4	0.2	0.039	2	
-->CLUSTER.4	7	0.714	0.25	0.06	2	
TOTAL	Elapsed	Time	for	CES+	:	483 ms
Fwithin	:	0.04096013				
Fbetween	:	117.216838	DefaultValues+content=0.4			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.053	1	
-->CLUSTER.2	10	0.5	0.333	0.022	3	
-->CLUSTER.3	8	0.875	0.35	0.037	2	
-->CLUSTER.4	2	1	0.286	0.031	7	
TOTAL	Elapsed	Time	for	CES+	:	433 ms
Fwithin	:	0.0355969				
Fbetween	:	180.557204	DefaultValues+content=0.6			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.077	0.061	1	
-->CLUSTER.2	10	0.7	0.467	0.014	3	
-->CLUSTER.3	5	0.6	0.15	0.014	2	
TOTAL	Elapsed	Time	for	CES+	:	365 ms
Fwithin	:	0.02960522				
Fbetween	:	194.914829	DefaultValues+content=0.7			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	5	0.4	0.077	0.053	1	
-->CLUSTER.2	12	0.75	0.6	0.041	3	
-->CLUSTER.3	11	0.727	0.4	0.044	2	
-->CLUSTER.4	9	0.778	0.583	0.05	4	
TOTAL	Elapsed	Time	for	CES+	:	637 ms
Fwithin	:	0.04702829				
Fbetween	:	45.9026026	DefaultValues+content=0.3			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	2	1	0.4	0.012	8	
-->CLUSTER.2	10	0.7	0.35	0.047	2	
-->CLUSTER.3	3	0.667	0.286	0.032	6	
-->CLUSTER.4	5	1	0.333	0.031	3	
-->CLUSTER.5	6	1	0.5	0.044	4	
TOTAL	Elapsed	Time	for	CES+	:	785 ms
Fwithin	:	0.03321605				
Fbetween	:	66.4168788	DefaultValues+content=0.2			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	11	0.545	0.3	0.055	2	
-->CLUSTER.2	4	1	0.267	0.031	3	
TOTAL	Elapsed	Time	for	CES+	:	779 ms
Fwithin	:	0.04325622				
Fbetween	:	37.1555969	DefaultValues+content=0.1			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	9	0.444	0.154	0.065	1	
-->CLUSTER.2	12	0.75	0.6	0.041	3	
-->CLUSTER.3	12	0.75	0.45	0.048	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES+	:	646 ms
Fwithin	:	0.05120367				
Fbetween	:	37.6689734	DefaultValues+content=0.3+tfidf=0.005			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	11	0.727	0.4	0.044	2	
-->CLUSTER.2	5	0.8	0.333	0.04	4	
TOTAL	Elapsed	Time	for	CES+	:	622 ms
Fwithin	:	0.04232717				
Fbetween	:	38.9908505	DefaultValues+content=0.3+tfidf=0.007			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	12	0.417	0.192	0.07	1	
-->CLUSTER.2	15	0.733	0.733	0.044	3	
-->CLUSTER.3	13	0.769	0.5	0.054	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES+	:	655 ms
Fwithin	:	0.05456437				
Fbetween	:	34.039564	DefaultValues+content=0.3+tfidf=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.731	0.163	1	
-->CLUSTER.2	16	0.688	0.733	0.045	3	
-->CLUSTER.3	15	0.8	0.6	0.059	2	
-->CLUSTER.4	12	0.583	0.583	0.051	4	
TOTAL	Elapsed	Time	for	CES+	:	669 ms
Fwithin	:	0.07940455				
Fbetween	:	25.8438552	DefaultValues+content=0.3+tfidf=0.003			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	29	0.655	0.731	0.148	1	
-->CLUSTER.2	19	0.632	0.6	0.067	2	
-->CLUSTER.3	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES+	:	455 ms
Fwithin	:	0.08879105				
Fbetween	:	25.0069267	DefaultValues+content=0.3+tfidf=0.003+CC=0.2			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	36	0.583	0.808	0.172	1	
-->CLUSTER.2	15	0.733	0.55	0.049	2	
-->CLUSTER.3	7	0.714	0.417	0.059	4	
-->CLUSTER.4	4	1	0.267	0.025	3	
TOTAL	Elapsed	Time	for	CES+	:	939 ms
Fwithin	:	0.07619204				
Fbetween	:	30.7665364	DefaultValues+content=0.3+tfidf=0.003+CC=0.4			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	14	0.357	0.192	0.071	1	
-->CLUSTER.2	17	0.529	0.45	0.059	2	
-->CLUSTER.3	12	0.75	0.6	0.038	3	
TOTAL	Elapsed	Time	for	CES+	:	1386 ms
Fwithin	:	0.05614523				
Fbetween	:	31.7058386	DefaultValues+content=0.3+tfidf=0.003+CC=0.5			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	13	0.385	0.192	0.068	1	
-->CLUSTER.2	16	0.5	0.4	0.06	2	
-->CLUSTER.3	5	1	0.333	0.028	3	
TOTAL	Elapsed	Time	for	CES+	:	1731 ms
Fwithin	:	0.05191907				
Fbetween	:	38.3944264	DefaultValues+content=0.3+tfidf=0.003+CC=0.6			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	14	0.357	0.192	0.071	1	
-->CLUSTER.2	13	0.538	0.35	0.053	2	
-->CLUSTER.3	5	1	0.333	0.028	3	
-->CLUSTER.4	6	1	0.5	0.039	4	
TOTAL	Elapsed	Time	for	CES+	:	1784 ms
Fwithin	:	0.04787114				
Fbetween	:	40.040978	DefaultValues+content=0.3+tfidf=0.003+CC=0.7			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	30	0.533	0.615	0.174	1	
-->CLUSTER.2	10	0.4	0.2	0.052	2	
-->CLUSTER.3	6	0.333	0.077	0.045	1	
-->CLUSTER.4	24	0.417	0.667	0.067	3	
-->CLUSTER.5	5	0.4	0.286	0.037	7	
TOTAL	Elapsed	Time	for	CES+	:	280 ms
Fwithin	:	0.07484956				
Fbetween	:	29.3105081	DefaultValues+content=0.3+tfidf=0.003+CC=0.1			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	21	0.381	0.308	0.091	1	
-->CLUSTER.2	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES+	:	616 ms
Fwithin	:	0.07141736				
Fbetween	:	27.2926539	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.02			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	25	0.32	0.308	0.089	1	
-->CLUSTER.2	23	0.565	0.867	0.051	3	
TOTAL	Elapsed	Time	for	CES+	:	686 ms
Fwithin	:	0.07024846				
Fbetween	:	24.4973725	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.001			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	26	0.769	0.769	0.154	1	
-->CLUSTER.2	14	0.786	0.55	0.064	2	
-->CLUSTER.3	2	0.5	0.05	0.018	2	
-->CLUSTER.4	24	0.542	0.867	0.053	3	
TOTAL	Elapsed	Time	for	CES+	:	410 ms
Fwithin	:	0.07219857				
Fbetween	:	32.5883863	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	0.792	0.731	0.152	1	
-->CLUSTER.2	14	0.786	0.55	0.064	2	
-->CLUSTER.3	2	0.5	0.05	0.018	2	
-->CLUSTER.4	24	0.542	0.867	0.053	3	
TOTAL	Elapsed	Time	for	CES+	:	410 ms
Fwithin	:	0.07172974				
Fbetween	:	32.6839483	DefaultValues+content=0.3+tfidf=0.003+CC=0.2+E=0.005			
DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.002 - Deadlock						
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	0.467	0.269	0.105	1	
-->CLUSTER.2	3	1	0.2	0.023	3	
TOTAL	Elapsed	Time	for	CES+	:	1559 ms
Fwithin	:	0.06381729				
Fbetween	:	45.2150538	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.001			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	28	0.679	0.731	0.151	1	
-->CLUSTER.2	19	0.737	0.7	0.061	2	
-->CLUSTER.3	13	0.769	0.833	0.067	4	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	6	0.833	0.333	0.032	3	
TOTAL	Elapsed	Time	for	CES+	:	410 ms
Fwithin	:	0.06959788				
Fbetween	:	31.7307039	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.004			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	1	0.577	0.152	1	
-->CLUSTER.2	24	0.625	0.75	0.067	2	
-->CLUSTER.3	13	0.923	1	0.071	4	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	4	1	0.267	0.025	3	
TOTAL	Elapsed	Time	for	CES+	:	410 ms
Fwithin	:	0.07028005				
Fbetween	:	36.5533538	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.005			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	15	1	0.577	0.152	1	
-->CLUSTER.2	16	0.625	0.5	0.06	2	
-->CLUSTER.3	3	0.333	0.05	0.032	2	
-->CLUSTER.4	4	0.75	0.429	0.037	6	
-->CLUSTER.5	4	1	0.267	0.025	3	
-->CLUSTER.6	10	1	0.833	0.056	4	
TOTAL	Elapsed	Time	for	CES+	:	743 ms
Fwithin	:	0.06020539				
Fbetween	:	44.7976622	DefaultValues+content=0.3+tfidf=0.003+CC=0.4+E=0.006			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	6	0.833	0.714	0.036	6	
-->CLUSTER.6	13	1	0.867	0.034	3	
-->CLUSTER.7	5	1	0.625	0.028	5	
TOTAL	Elapsed	Time	for	CES+	:	792 ms
Fwithin	:	0.06378233				
Fbetween	:	39.9852981	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0001			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	3	1	0.115	0.163	1	
-->CLUSTER.2	9	0.556	0.25	0.041	2	
TOTAL	Elapsed	Time	for	CES+	:	737 ms
Fwithin	:	0.10208461				
Fbetween	:	77.4258518	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.003			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	14	1	0.933	0.033	3	
-->CLUSTER.7	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES+	:	806 ms
Fwithin	:	0.06327724				
Fbetween	:	39.9750019	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0009			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	22	1	0.846	0.155	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	14	1	0.933	0.033	3	
-->CLUSTER.7	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES+	:	794 ms
Fwithin	:	0.06327724				
Fbetween	:	39.9750019	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0008			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	1	0.923	0.168	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	2	0.5	0.05	0.017	2	
-->CLUSTER.7	14	1	0.933	0.033	3	
-->CLUSTER.8	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES+	:	801 ms
Fwithin	:	0.0590203				
Fbetween	:	46.6146868	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0005			

Table 32: Best results on ParallelDS with optimum parameter values for CES+

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	24	1	0.923	0.168	1	
-->CLUSTER.2	26	0.654	0.85	0.069	2	
-->CLUSTER.3	4	0.5	0.286	0.054	7	
-->CLUSTER.4	13	0.923	1	0.071	4	
-->CLUSTER.5	7	0.857	0.857	0.034	6	
-->CLUSTER.6	2	0.5	0.05	0.017	2	
-->CLUSTER.7	14	1	0.933	0.033	3	
-->CLUSTER.8	6	1	0.75	0.027	5	
TOTAL	Elapsed	Time	for	CES+	:	798 ms
Fwithin	:	0.0590203				
Fbetween	:	46.6146868	content=0.3+tfidf=0.003+CC=0.4+E=0.005+S=0.0001			

1.8.6-3 GoDaddyDS

GoDaddyDS						
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	10	0.9	0.22	0.129	1	
-->CLUSTER.2	34	0.588	0.488	0.157	1	
-->CLUSTER.3	5	0.8	0.16	0.065	3	
-->CLUSTER.4	9	1	0.36	0.064	3	
TOTAL	Elapsed	Time	for	CES+	:	596 ms
Fwithin	:	0.10367349				
Fbetween	:	47.237956	DefaultValues.txt			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.463	0.146	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	13	1	0.52	0.087	3	
TOTAL	Elapsed	Time	for	CES+	:	736 ms
Fwithin	:	0.09222384				
Fbetween	:	48.7220668	DefaultValues+content=0.4			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	31	0.548	0.415	0.182	1	
-->CLUSTER.2	15	1	0.6	0.096	3	
-->CLUSTER.3	3	0.333	0.024	0.051	1	
TOTAL	Elapsed	Time	for	CES+	:	662 ms
Fwithin	:	0.10999654				
Fbetween	:	55.4317021	DefaultValues+content=0.6			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	34	0.559	0.463	0.128	1	
-->CLUSTER.2	3	0.333	0.024	0.04	1	
-->CLUSTER.3	13	1	0.52	0.082	3	
TOTAL	Elapsed	Time	for	CES+	:	771 ms
Fwithin	:	0.08330776				
Fbetween	:	44.422459	DefaultValues+content=0.3			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	15	1	0.6	0.088	3	
TOTAL	Elapsed	Time	for	CES+	:	784 ms
Fwithin	:	0.09731533				
Fbetween	:	41.6826159	DefaultValues+content=0.4+tfidf=0.005			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	31	0.548	0.415	0.146	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	13	1	0.52	0.087	3	
TOTAL	Elapsed	Time	for	CES+	:	754 ms
Fwithin	:	0.09208441				
Fbetween	:	49.5678461	DefaultValues+content=0.4+tfidf=0.007			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	16	1	0.64	0.092	3	
TOTAL	Elapsed	Time	for	CES+	:	784 ms
Fwithin	:	0.09853628				
Fbetween	:	41.2808373	DefaultValues+content=0.4+tfidf=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.596	0.756	0.163	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	
-->CLUSTER.3	16	1	0.64	0.092	3	
TOTAL	Elapsed	Time	for	CES+	:	800 ms
Fwithin	:	0.09945965				
Fbetween	:	41.1223687	DefaultValues+content=0.4+tfidf=0.003			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	50	0.58	0.707	0.16	1	
-->CLUSTER.2	3	0.333	0.024	0.044	1	

-->CLUSTER.3	15	1	0.6	0.088	3	
TOTAL	Elapsed	Time	for	CES+	:	799 ms
Fwithin	:	0.09731533				
Fbetween	:	41.6826159	DefaultValues+content=0.4+tfidf=0.0045			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	46	0.63	0.707	0.159	1	
-->CLUSTER.2	16	0.938	0.6	0.09	3	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
TOTAL	Elapsed	Time	for	CES+	:	519 ms
Fwithin	:	0.0943803				
Fbetween	:	47.9574899	DefaultValues+content=0.4+tfidf=0.004+CC=0.2			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	42	0.714	0.882	0.154	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	12	1	0.48	0.081	3	
TOTAL	Elapsed	Time	for	CES+	:	961 ms
Fwithin	:	0.11595811				
Fbetween	:	37.9654753	DefaultValues+content=0.4+tfidf=0.004+CC=0.4			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	32	0.938	0.732	0.14	1	
-->CLUSTER.2	11	0.636	0.28	0.082	3	
-->CLUSTER.3	29	0.862	0.735	0.107	2	
-->CLUSTER.4	9	1	0.36	0.067	3	
TOTAL	Elapsed	Time	for	CES+	:	1186 ms
Fwithin	:	0.09906089				
Fbetween	:	37.5409833	DefaultValues+content=0.4+tfidf=0.004+CC=0.5			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.842	0.78	0.151	1	
-->CLUSTER.2	23	0.826	0.76	0.102	3	
-->CLUSTER.3	5	0.6	0.073	0.059	1	
-->CLUSTER.4	16	1	0.471	0.076	2	
TOTAL	Elapsed	Time	for	CES+	:	1588 ms
Fwithin	:	0.09712504				
Fbetween	:	40.0123355	DefaultValues+content=0.4+tfidf=0.004+CC=0.6			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	32	0.719	0.561	0.138	1	
-->CLUSTER.2	17	0.765	0.52	0.096	3	
-->CLUSTER.3	2	0.5	0.024	0.018	1	
-->CLUSTER.4	16	1	0.471	0.078	2	
TOTAL	Elapsed	Time	for	CES+	:	2412 ms
Fwithin	:	0.08248866				
Fbetween	:	43.1753983	DefaultValues+content=0.4+tfidf=0.004+CC=0.7			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	40	0.725	0.707	0.149	1	
-->CLUSTER.2	16	0.625	0.294	0.096	2	
-->CLUSTER.3	5	1	0.147	0.043	2	
TOTAL	Elapsed	Time	for	CES+	:	2819 ms
Fwithin	:	0.09595713				
Fbetween	:	25.6707884	DefaultValues+content=0.4+tfidf=0.004+CC=0.8			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	54	0.593	0.78	0.167	1	
-->CLUSTER.2	4	0.5	0.049	0.053	1	
-->CLUSTER.3	5	1	0.2	0.04	3	
TOTAL	Elapsed	Time	for	CES+	:	471 ms
Fwithin	:	0.08639169				
Fbetween	:	47.5998202	DefaultValues+content=0.4+tfidf=0.004+CC=0.1			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.596	0.756	0.167	1	
-->CLUSTER.2	15	0.933	0.56	0.09	3	
TOTAL	Elapsed	Time	for	CES+	:	650 ms
Fwithin	:	0.12844943				
Fbetween	:	33.4529307	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.002			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	42	0.619	0.634	0.156	1	
-->CLUSTER.2	8	0.5	0.098	0.073	1	
-->CLUSTER.3	16	0.938	0.6	0.09	3	
-->CLUSTER.4	2	0.5	0.024	0.034	1	
TOTAL	Elapsed	Time	for	CES+	:	514 ms
Fwithin	:	0.08836014				
Fbetween	:	49.0565863	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.526	0.488	0.15	1	
-->CLUSTER.2	15	0.933	0.56	0.09	3	
TOTAL	Elapsed	Time	for	CES+	:	828 ms
Fwithin	:	0.11992432				
Fbetween	:	39.070154	DefaultValues+content=0.4+tfidf=0.004+CC=0.2+E=0.001			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	28	0.536	0.441	0.146	2	
-->CLUSTER.2	25	0.92	0.561	0.115	1	
-->CLUSTER.3	13	1	0.52	0.085	3	
TOTAL	Elapsed	Time	for	CES+	:	978 ms
Fwithin	:	0.1153912				
Fbetween	:	40.4554957	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.002			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	37	0.703	0.765	0.149	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
-->CLUSTER.4	13	1	0.52	0.085	3	
TOTAL	Elapsed	Time	for	CES+	:	514 ms
Fwithin	:	0.09523945				
Fbetween	:	49.0641692	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.004			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	52	0.538	0.683	0.165	1	
-->CLUSTER.2	10	0.7	0.171	0.093	1	
-->CLUSTER.3	14	1	0.56	0.08	3	
TOTAL	Elapsed	Time	for	CES+	:	1228 ms
Fwithin	:	0.11271651				
Fbetween	:	37.6016151	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.001			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	38	0.684	0.765	0.152	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	2	0.5	0.024	0.034	1	
-->CLUSTER.4	12	1	0.48	0.081	3	
TOTAL	Elapsed	Time	for	CES+	:	923 ms
Fwithin	:	0.0951172				
Fbetween	:	48.8829746	DefaultValues+content=0.4+tfidf=0.004+CC=0.4+E=0.0035			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
TOTAL	Elapsed	Time	for	CES+	:	939 ms
Fwithin	:	0.11810412				
Fbetween	:	37.5378643	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.001			

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	36	0.694	0.735	0.139	2	
-->CLUSTER.2	21	0.905	0.463	0.108	1	
-->CLUSTER.3	6	1	0.24	0.07	3	
TOTAL	Elapsed	Time	for	CES+	:	939 ms
Fwithin	:	0.10533146				
Fbetween	:	36.0946898	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.003			
CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
-->CLUSTER.4	2	1	0.08	0.012	3	
TOTAL	Elapsed	Time	for	CES+	:	978 ms
Fwithin	:	0.09167867				
Fbetween	:	45.957703	content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.0009			

Table 33: Best results on GoDaddyDS with the optimum parameter values for CES+

CES+	Size	Precision	Recall	StandartD.	Considered	
-->CLUSTER.1	44	0.727	0.941	0.155	2	
-->CLUSTER.2	25	0.92	0.561	0.113	1	
-->CLUSTER.3	17	1	0.68	0.087	3	
-->CLUSTER.4	2	1	0.08	0.012	3	
TOTAL Elapsed Time for CES+ : 964 ms						
Fwithin : 0.0916786674710695			content=0.4+tfidf=0.004+CC=0.4+E=0.003+S=0.0008			
Fbetween : 45.9577029848812						

APPENDIX SECTION 2 – MOGA GENERATION NUMBER RESUTLS

2.1 PrincetonDS

Number of Generations	Best Pareto Individual	FWithin	FBetween/1000
1	1.Individual	0.09	0.06842
25	2.Individual	0.08	0.10264
50	3.Individual	0.09	0.09028
100	4.Individual	0.08	0.11527
150	5.Individual	0.07	0.11703
200	6.Individual	0.06	0.1403
250	7.Individual	0.06	0.14762
300	8.Individual	0.07	0.19609
350	9.Individual	0.06	0.19609
400	10.Individual	0.06	0.19609
450	11.Individual	0.07	0.19609
500	12.Individual	0.06	0.19118
550	13.Individual	0.06	0.19118
600	14.Individual	0.06	0.16206
650	15.Individual	0.06	0.19118
700	16.Individual	0.06	0.18538
750	17.Individual	0.07	0.19681
800	18.Individual	0.07	0.19681
850	19.Individual	0.07	0.15152
900	20.Individual	0.07	0.18921
950	21.Individual	0.07	0.1721
1000	22.Individual	0.07	0.18921

2.2 ParallelsDS

Number of Generations	Best Pareto Individual	Fwithin	FBetween/1000
1	1.Individual	0.05	0.07477
25	2.Individual	0.05	0.08516
50	3.Individual	0.05	0.09432
100	4.Individual	0.05	0.10563
150	5.Individual	0.05	0.13235
200	6.Individual	0.05	0.15617
250	7.Individual	0.05	0.17388
300	8.Individual	0.05	0.17493
350	9.Individual	0.05	0.17493
400	10.Individual	0.05	0.18093
450	11.Individual	0.05	0.15985
500	12.Individual	0.05	0.15985
550	13.Individual	0.06	0.18564
600	14.Individual	0.06	0.20954
650	15.Individual	0.05	0.20954
700	16.Individual	0.05	0.20954
750	17.Individual	0.05	0.17388
800	18.Individual	0.05	0.15617
850	19.Individual	0.05	0.14987
900	20.Individual	0.05	0.16636
950	21.Individual	0.05	0.17388
1000	22.Individual	0.05	0.17388

2.3 GoDaddyDS

Number of Generations	Best Pareto Individual	FWithin	FBetween/1000
1	1.Individual	0.1	0.04484
25	2.Individual	0.08	0.07451
50	3.Individual	0.06	0.11037
100	4.Individual	0.06	0.11379
150	5.Individual	0.06	0.10977
200	6.Individual	0.05	0.11037
250	7.Individual	0.06	0.11379
300	8.Individual	0.05	0.11379
350	9.Individual	0.05	0.12201
400	10.Individual	0.05	0.12062
450	11.Individual	0.05	0.12201
500	12.Individual	0.05	0.1325
550	13.Individual	0.05	0.12201
600	14.Individual	0.05	0.12061
650	15.Individual	0.05	0.12062
700	16.Individual	0.05	0.11379
750	17.Individual	0.05	0.11379
800	18.Individual	0.05	0.11379
850	19.Individual	0.05	0.11379
900	20.Individual	0.05	0.11037
950	21.Individual	0.05	0.10977
1000	22.Individual	0.05	0.11379

APPENDIX SECTION 3- CES+MOGA RESULTS

3.1 PrincetonDS

Part1	***Test1***	***Test2***	***Test3***	***Test4***	***Test5***
Fwithin :	0.227	0.239	0.256	0.256	0.245
Fbetween :	33.446	24.524	29.974	29.974	26.508
AvgPrecision:	0.444	0.417	0.417	0.417	0.417
AvgRecall :	0.583	0.750	0.750	0.750	0.750
F1 :	0.505	0.536	0.536	0.536	0.536
Part2					
Fwithin :	0.205	0.205	0.294	0.294	0.223
Fbetween :	42.470	42.470	62.459	62.459	47.883
AvgPrecision:	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.889	0.889	0.889	0.889	0.889
F1 :	0.941	0.941	0.941	0.941	0.941
Part3					
Fwithin :	0.330	0.269	0.243	0.263	0.225
Fbetween :	50.992	39.979	42.202	44.965	31.105
AvgPrecision:	0.889	0.889	1.000	1.000	0.889
AvgRecall :	0.889	0.889	1.000	1.000	0.889
F1 :	0.889	0.889	1.000	1.000	0.889
Part4					
Fwithin :	0.214	0.199	0.209	0.209	0.199
Fbetween :	36.617	27.097	33.700	33.700	27.097
AvgPrecision:	0.722	0.722	0.722	0.722	0.722
AvgRecall :	0.833	0.833	0.833	0.833	0.833
F1 :	0.774	0.774	0.774	0.774	0.774
Part5					
Fwithin :	0.271	0.242	0.222	0.249	0.242
Fbetween :	43.085	47.308	40.225	35.259	32.325
AvgPrecision:	0.889	1.000	1.000	0.889	0.889
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	0.941	1.000	1.000	0.941	0.941
Part6					
Fwithin :	0.349	0.300	0.349	0.325	0.300
Fbetween :	43.328	39.391	43.328	41.514	39.391
AvgPrecision:	0.750	0.750	0.750	0.750	0.750
AvgRecall :	0.750	0.750	0.750	0.750	0.750
F1 :	0.750	0.750	0.750	0.750	0.750

Part7					
Fwithin :	0.200	0.174	0.207	0.266	0.205
Fbetween :	36.904	57.041	47.800	40.273	44.587
AvgPrecision:	0.833	0.750	0.833	0.722	0.833
AvgRecall :	0.750	0.625	0.750	0.833	0.750
F1 :	0.789	0.682	0.789	0.774	0.789
Part8					
Fwithin :	0.198	0.181	0.198	0.202	0.185
Fbetween :	59.152	39.350	59.152	63.131	44.977
AvgPrecision:	0.889	0.889	0.889	0.889	0.889
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	0.941	0.941	0.941	0.941	0.941
Part9					
Fwithin :	0.206	0.200	0.194	0.211	0.188
Fbetween :	35.779	33.178	30.317	38.131	27.194
AvgPrecision:	0.778	0.778	0.778	0.778	0.778
AvgRecall :	0.889	0.889	0.889	0.889	0.889
F1 :	0.830	0.830	0.830	0.830	0.830
Part10					
Fwithin :	0.233	0.243	0.233	0.213	0.213
Fbetween :	25.345	25.863	25.345	23.682	23.682
AvgPrecision:	0.750	0.750	0.750	0.750	0.750
AvgRecall :	0.833	0.833	0.833	0.833	0.833
F1 :	0.789	0.789	0.789	0.789	0.789

3.2 ParallelsDS

Part1	***Test1***	***Test2***	***Test3***	***Test4***	***Test5***
Fwithin :	0.248	0.177	0.299	0.192	0.275
Fbetween :	51.930	46.121	54.999	48.579	36.679
AvgPrecision :	1.000	1.000	1.000	1.000	0.875
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	1.000	1.000	1.000	1.000	0.933
Part2					
Fwithin :	0.385	0.326	0.274	0.300	0.326
Fbetween :	16.867	21.968	21.427	21.739	21.968
AvgPrecision :	0.800	0.875	0.875	0.875	0.875
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	0.889	0.933	0.933	0.933	0.933
Part3					
Fwithin :	0.000	0.000	0.425	0.000	0.382
Fbetween :	NaN	NaN	51.845	NaN	51.291
AvgPrecision :	1.000	1.000	0.750	1.000	0.750
AvgRecall :	0.800	0.800	0.300	0.800	0.300
F1 :	0.889	0.889	0.429	0.889	0.429
Part4					
Fwithin :	0.064	0.064	0.064	0.064	0.064
Fbetween :	38.596	38.596	38.596	38.596	38.596
AvgPrecision :	1.000	1.000	1.000	1.000	1.000
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	1.000	1.000	1.000	1.000	1.000
Part5					
Fwithin :	0.206	0.235	0.281	0.198	0.198
Fbetween :	25.920	23.409	24.079	30.351	30.351
AvgPrecision :	0.917	0.750	0.750	0.722	0.722
AvgRecall :	1.000	0.833	1.000	0.778	0.778
F1 :	0.957	0.789	0.857	0.749	0.749
Part6					
Fwithin :	0.000	0.096	0.096	0.096	0.109
Fbetween :	NaN	85.704	85.704	85.704	63.711
AvgPrecision :	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.500	0.750	0.750	0.750	0.750
F1 :	0.667	0.857	0.857	0.857	0.857

Part7					
Fwithin :	0.081	0.143	0.081	0.097	0.097
Fbetween :	40.061	31.577	40.061	37.335	37.335
AvgPrecision :	1.000	1.000	1.000	1.000	1.000
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	1.000	1.000	1.000	1.000	1.000
Part8					
Fwithin :	0.078	0.078	0.078	0.080	0.078
Fbetween :	28.686	28.686	28.686	26.072	28.686
AvgPrecision :	0.875	0.875	0.875	0.875	0.875
AvgRecall :	0.750	0.750	0.750	0.750	0.750
F1 :	0.808	0.808	0.808	0.808	0.808
Part9					
Fwithin :	0.189	0.136	0.189	0.185	0.185
Fbetween :	17.676	24.995	17.676	16.796	16.796
AvgPrecision :	0.875	0.833	0.875	0.875	0.875
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	0.933	0.909	0.933	0.933	0.933
Part10					
Fwithin :	0.206	0.206	0.206	0.206	0.206
Fbetween :	16.744	16.744	16.744	16.744	16.744
AvgPrecision :	0.867	0.867	0.867	0.867	0.867
AvgRecall :	1.000	1.000	1.000	1.000	1.000
F1 :	0.929	0.929	0.929	0.929	0.929

3.3 GoDaddyDS

Part1	***Test1***	***Test2***	***Test3***	***Test4***	***Test5***
Fwithin :	0.178	0.194	0.161	0.161	0.161
Fbetween :	32.729	24.811	28.620	28.620	28.620
AvgPrecision:	0.500	0.417	0.500	0.500	0.833
AvgRecall :	0.611	0.750	0.611	0.611	0.667
F1 :	0.550	0.536	0.550	0.550	0.741
Part2					
Fwithin :	0.253	0.229	0.252	0.252	0.253
Fbetween :	36.444	33.849	29.366	29.366	36.444
AvgPrecision:	1.000	1.000	0.867	0.867	1.000
AvgRecall :	1.000	1.000	0.833	0.833	1.000
F1 :	1.000	1.000	0.850	0.850	1.000
Part3					
Fwithin :	0.208	0.164	0.107	0.208	0.120
Fbetween :	31.284	21.096	58.954	31.284	62.115
AvgPrecision:	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.857	0.857	0.524	0.857	0.524
F1 :	0.923	0.923	0.688	0.923	0.688
Part4					
Fwithin :	0.129	0.173	0.151	0.129	0.283
Fbetween :	34.438	46.268	41.471	34.438	29.936
AvgPrecision:	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.550	0.550	0.550	0.550	0.800
F1 :	0.710	0.710	0.710	0.710	0.889
Part5					
Fwithin :	0.192	0.205	0.187	0.187	0.205
Fbetween :	23.676	34.838	14.919	14.919	34.838
AvgPrecision:	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.675	0.675	0.775	0.775	0.675
F1 :	0.806	0.806	0.873	0.873	0.806
Part6					
Fwithin :	0.000	0.000	0.000	0.000	0.000
Fbetween :	NaN	NaN	NaN	NaN	NaN
AvgPrecision:	1.000	1.000	1.000	1.000	1.000
AvgRecall :	0.286	0.286	0.714	0.286	0.286
F1 :	0.444	0.444	0.833	0.444	0.444

Part7					
Fwithin :	0.000	0.207	0.207	0.207	0.173
Fbetween :	NaN	12.521	12.521	12.521	26.030
AvgPrecision:	0.800	0.900	0.900	0.900	0.889
AvgRecall :	0.800	0.775	0.775	0.775	0.517
F1 :	0.800	0.833	0.833	0.833	0.653
Part8					
Fwithin :	0.190	0.259	0.190	0.259	0.259
Fbetween :	35.436	41.403	35.436	41.403	41.403
AvgPrecision:	0.833	0.833	0.833	0.833	0.833
AvgRecall :	0.689	0.689	0.689	0.689	0.689
F1 :	0.754	0.754	0.754	0.754	0.754
Part9					
Fwithin :	0.120	0.236	0.120	0.120	0.120
Fbetween :	58.593	83.757	58.593	58.593	58.593
AvgPrecision:	0.889	0.889	0.889	0.889	0.889
AvgRecall :	0.722	0.722	0.722	0.722	0.722
F1 :	0.797	0.797	0.797	0.797	0.797
Part10					
Fwithin :	0.128	0.128	0.198	0.198	0.198
Fbetween :	31.445	31.445	40.799	40.799	40.799
AvgPrecision:	1.000	1.000	0.917	0.917	0.917
AvgRecall :	0.700	0.700	0.800	0.800	0.800
F1 :	0.824	0.824	0.854	0.854	0.854