Mastering Data-Intensive Collaboration and Decision Making

FP7 - Information and Communication Technologies

Grant Agreement no: 257184
Collaborative Project
Project start: 1 September 2010, Duration: 36 months

D4.2.1 – The Dicode Decision Making Support Services (initial version)

Due date of deliverable:	31 August 2011
Actual submission date:	31 August 2011
Responsible Partner:	CTI
Contributing Partners:	CTI, FHG, UOL

Nature: ☒ Report ☒ Prototype ☐ Demonstrator ☐ Other

Dissemination Level:

- ☒ PU : Public
- ☐ PP : Restricted to other programme participants (including the Commission Services)
- ☐ RE : Restricted to a group specified by the consortium (including the Commission Services)
- ☐ CO : Confidential, only for members of the consortium (including the Commission Services)

Keyword List: Decision making support services, recommendation mechanism, semantic-driven collaboration monitoring mechanism, decision making support, abstract service description, interfaces, operations, development

The Dicode project (dicode-project.eu) is funded by the European Commission, Information Society and Media Directorate General, under the FP7 Cooperation programme (ICT/SO4.3: Intelligent Information Management).

The Dicode Consortium

Research Academic Computer Technology Institute (CTI) (coordinator), Greece

University of Leeds(UOL), UK

Fraunhofer-Gesellschaft zur Foerderung der angewandten Forschung e.V. (FHG), Germany

Universidad Politécnica De Madrid(UPM), Spain

Neofonie GmbH(NEO), Germany

Image Analysis Limited(IMA), UK

Biomedical Research Foundation, Academy of Athens(BRF), Greece

Publicis Frankfurt Zweigniederlassung der PWW GmbH (PUB), Germany

Document history			
Version	Date	Status	Modifications made by
1	30-07-2011	First draft version	Manolis Tzagarakis (CTI)
2	15-08-2011	Second draft version	Axel Poigné (FHG) Ahmad Ammari (UOL) Manolis Tzagarakis (CTI)
3	20-08-2011	Third draft version with formatting modifications (sent to internal reviewers)	Manolis Tzagarakis (CTI)
4	26-08-2011	Fourth version; reviewer's comments incorporated (sent to SC)	Axel Poigné (FHG) Ahmad Ammari (UOL) Manolis Tzagarakis (CTI)
5	31-08-2011	Final version (approved by SC, sent to the Project Officer)	Manolis Tzagarakis (CTI)

Deliverable managers

- Manolis Tzagarakis, CTI

List of Contributors

- Manolis Tzagarakis, CTI
- Spyros Christodoulou, CTI
- Nikos Karacapilidis, CTI
- Axel Poigné, FHG
- Ahmad Ammari, UOL

List of Evaluators

- Stephan Rueping, FHG
- Guillermo de la Calle Velasco, UPM

Summary

This deliverable is to be considered as a progress report on the initial version of the Dicode decision making support services, which are designed and implemented in the context of WP4. Decision making support services developed in Dicode concern the building of machine-interpretable knowledge in order to actively support various decision making tasks. In this deliverable, the technical specifications of the decision making support services related to Tasks 4.2, 4.3 and 4.5 of WP4 are presented. The intended audience of this document are designers and developers of the Dicode project. The document informs them on which decision making support services have been developed and how they can be used, thus providing a first framework for discussing how to proceed with the full development of the envisioned services. The initial versions of the services are presented using a formalized and project-wide adopted service description template.

Table of Contents

1 Introduction

1.1 Context

This deliverable presents the initial version of the decision making support services that have been designed and developed in the context of WP4 ("Data-Intensive Collaboration & Decision Support Services") of the Dicode project. More specifically, it reports on the progress of work being carried out, by describing the related initial version of services that have been developed, in Tasks 4.2, 4.3 and 4.5.

This is the first of a series of deliverables reporting on the progress of work with respect to the decision making support services in the context of WP4. While the focus of this deliverable (D4.2.1) is on the initial version of collaboration support services, deliverable D4.2.2 (due in month 24) will report on their enhanced version. The final version of the decision making support services will be reported in deliverable D4.2.3 (due in month 33).

1.2 Objectives

The purpose of this document is to present the initial version of the developed services related to decision making support in the context of Tasks 4.2, 4.3 and 4.5, as they originated from the functional specifications outlined in deliverable D2.2 in order to frame and start the implementation related discussion on how to realize the vision of the Dicode project.

The developed services are presented from a technical perspective, broken down to the level of individual operations, in order to show their role and use and facilitate their assessment with respect to the derived functional specifications. The operations presented are those which are available to clients to be invoked and executed without going into detail about how exactly these can be invoked or executed. In particular, the presented operations can be executed by various technologies such as REST (Fielding, 2000) or Web Services (W3C - Web Services Architecture, 2004), but such issues are not the focus of their description. The description of services takes an operation-oriented approach listing the available operations and detailing their aim and purpose.

The initial version of the services are presented using a service description template, which has been derived and used in the context of deliverable D3.1.1 ("The Dicode Data Mining Framework"), called the Abstract Service Description. The Abstract Service Description template provides a technical specification of services by providing an overview of the supported interfaces and the relevant operations. For each operation, a description along with major input and output information is presented.

2 Dicode Decision Making Support Services

The aim of the decision making support services is to turn information and knowledge machine interpretable in order to allow active participation of the system in collaborative activities, hence aiding the overall decision making process. In the context of Dicode, three different strategies are used to generate knowledge for the decision making process:

- Automatically generating knowledge about the preferences of a single user, which can for example be used to recommend documents he might be interested in.

- Automatically generating knowledge about the interaction of a whole group of users, e.g. by structuring a written discussion and identify related discussions.
- Allowing users to manually add knowledge to the system, which will allow to fully capture and describe all knowledge that is relevant in a decision making process.

Towards these overall goals, a number of relevant services will be developed that can fully address the relevant requirements of Dicode as outlined in deliverable D2.2.
This documents reports on the initial version of the services that have been developed in the context of the following tasks:

- Task 4.2 - Recommendation mechanism, which is concerned with the design and development of recommender mechanisms to support the user in the analysis of large and heterogeneous data;
- Task 4.3 - Semantic-driven collaboration monitoring mechanism, which will derive a semantic-driven model of collaborative processes in a community;
- Task 4.5 - Decision making support services, which is concerned with the formalization of collaborative decision making issues to intelligently support stakeholders in such activities.

In the following, we present the service description of the initial version of the services being developed in the abovementioned tasks of WP4. For each service, a short description related to its aim and purpose is given, followed by the abstract service description template of its initial version.

2.1 Recommendation mechanism

Recommendation mechanism is the focus of Task 4.2. In general, recommender systems attempt to recommend information items that are likely to be of interest to the user. "Typically, a recommender system compares a user profile[1] to some reference characteristics, and seeks to predict the 'rating' or 'preference' that a user would give to an item they had not yet considered"(Bell et al., 2007). Many algorithms have been used in recommender systems. These are often based on distance measures. The distance indicates the similarity of information items. Then, those items are recommended that are "closest" to match the user profile. Typically, fixed distance functions are used such as Euclidian distance[2] or cosine distance[3].

The key idea of similarity learning is to replace fixed distance functions by learning a function that produces a non-negative real number for any pair of examples. The intended semantic is that the higher this number the more similar the two examples are. The training data that the function learns from consist of example pairs labelled as similar or dissimilar (see Deliverable D3.1.1, Section 3.2.10). We also refer to Deliverable D3.2.1 regarding experiments and a prototypical evaluation for the recommendation mechanism.

2.1.1 Recommender Service

The Dicode Recommender Service (DRS) will be based on models generated by similarity learning. These models depend on the type of information items considered in Dicode.

[1] http://en.wikipedia.org/wiki/User_profile
[2] http://en.wikipedia.org/wiki/Euclidean_distance
[3] http://en.wikipedia.org/wiki/Cosine_similarity

Information items to be considered at present comprise user profiles, documents, and data sets. For every type, one or more specific models will be generated using the Dicode Similarity Learning Service (see Deliverable D3.2.1).

User interaction will be used as feedback to improve the similarity models in that the Dicode system will observe the choices made by the user with regard to the "similar items" offered, extending the list of similar items as input for similarity learning (see Deliverable D3.2.1).

DRS is generic, in that it will be parameterised by the type of information items, the similarity model, and output options.

The abstract service description of the initial version of DRS is presented below.

Name	*Recommender Service*
Standards	• OGC 05-008c1 Web Services Common Specification V1.0.
Description	The Recommender Service allows querying the Dicode system for information items that are similar to those provided as input. The search is based on similarity models learned by using the Dicode Similarity Learning Service. The Recommender Service provides its functionality through the following interfaces: • *ServiceCapabilities*: Informs about the common and specific capabilities. • *Recommendation*: provides query and feedback operations.
Interface	*ServiceCapabilities*
getCapabilities	Informs the requestor about the common and specific capabilities. Examples of specific capabilities are the supported information item types, similarity models related to the type, and output options. The *getCapabilities* operation of the Service Capabilities Interface is backward compatible with the concepts and definitions of the *getCapabilities* operation as defined in OGC 05-008c1 Web Services Common Specification V1.0.
Interface	*Recommendation*
query	Defines a query. Parameters include the type of information items, the similarity model, and the information item(s) for which similar items to be recommended. An optional integer parameter restricts the number of recommended items. The default is that all similar items are listed.
feedback	Provides a list of information item pairs considered to be similar by the user.

Example usage	A user may want to find similar data sets, for instance of clinical trials, or want to find users with a similar profile as the own profile.
Comments	• Since (at present) detailed specification of information types are not available yet, similarity models for the video lecture completion for recommender systems (http://tunedit.org/challenge/VLNetChallenge) have been generated as a proof-of-concept. • Computation times depend more or less linearly on the number of attributes and on the size of the data collection. • Improvement of results depends on improvement of similarity learning.
Conformance classes	Not available.
Implementation rules	Not available.
Implementation status	A prototypical version of the algorithm has been implemented. The algorithm is not yet provided as a service. For a snapshot of output data see Deliverable D3.2.1.
UML model	Not available.

2.2 Semantic-driven collaboration monitoring mechanism

The Community Modelling and User Profiling services comprise a set of services that address the challenges of the semantic-driven collaboration monitoring mechanism task (Task 4.3) of WP4.

Subsequently, an abstract, coarse description of the Community Modelling and User Profiling Services is provided in textual format according to Dicode Service Description Template. The description presents the basic functionality (in terms of interfaces and operations) of each of the services whose prototypical version has been implemented. These "abstract service descriptions" are meant as a high level introduction to the Community Modelling and User Profiling Services and will be complemented by formal abstract specifications and/or implementation specifications when the services become operational.

The abstract service description for each of the services is complemented by an appendix section that describes the experiments that have been carried out in order to demonstrate the prototypical version of the service. Each complementing appendix section includes detailed descriptions of the inputs, outputs, and methodology of operations that have been implemented for the current prototypical version of the service. Links to download the input and output files for each service are provided in the appendix section for that service as well.

2.2.1 Cluster Discussion Threads

In Section 7.6 of deliverable D2.2 - The Dicode Approach, the community modelling and user profiling services have been introduced to support Dicode recommendation services. These recommendation services will be able to use the output of the community modelling and user profiling services to recommend items (e.g. data set, a set of discussions, a topic of interests) to the user within Dicode communities. The "Cluster Discussion Threads" service is aimed to support discussion recommendations by taking the community discussions that exist in the Discussion-forum view of the Dicode collaboration space (See Section 7.4 of deliverable D2.2) as input and group them into distinct clusters based on their textual content. The service also produces the centroid of each cluster, which is a vector of the average term weights that exist in each cluster. The output of the service (assignment of discussions to clusters and cluster centroids) represents the foundation for a content-based recommendation service that compares the user characteristics with cluster characteristics based on the cluster centroids and then recommends to the user the discussions that belong to the cluster whose characteristics are the closest to the user characteristics.

The following abstract description of the service is given:

Name	Cluster Discussion Threads
Standards	
Description	The Cluster Discussion Threads service takes a collection of discussion threads and clusters the threads into groups, where each group contains a subset of the original threads. The discussion threads that belong to a particular group are relatively similar in content. The Cluster Discussion Threads service provides its functionality through the ThreadsClusterer interface. The purpose of this interface is to receive a collection of discussion threads as input, perform text pre-processing on the received threads, build and train a clustering model, generate clusters of discussion threads, assign each of the discussion threads to one of the generated clusters, and retrieve the centroid vector for each generated cluster.
Interface	ThreadsClusterer
readDiscussions	The *readDiscussions* operation reads the dataset from a storage resource. In the prototypical version of the service, the storage resource is an Excel sheet containing the required attributes of the dataset. In subsequent versions, the storage resource will be table / view in a relational database server.
dataToDocuments	The *dataToDocuments* operation transforms each discussion thread within the discussions dataset to a document. This step is necessary to treat each title as a separate instance (observation)
processDocuments	The *processDocuments* operation performs text pre-processing on the threads in the collection. The *processDocuments* operations include the

	following text pre-processing operators: tokenize documents, transform tokens to lower cases, filter stop words, and filter tokens by character length (See Section A.3.1 of Appendix A for more details about these operators). The output is a pre-processed version of the discussion threads
trainModel	The *trainModel* operation receives the pre-processed version of the discussion threads and uses them to build and train a number of discussion threads clustering models. The output is a number of trained clustering models, each model having a unique number of clusters of discussion threads.
validateModel	The *validateModel* operation validates the clustering models generated by the *trainModel* operation by calculating cluster validity measurements and selects the model having the best measurements as input to the *clusterThreads* operation. In the prototypical version of the service, this operation is represented by the 'cluster distance performance' and the 'item distribution performance' operators (See Section A.3.3 of Appendix A for more details about the operators).
clusterThreads	The *clusterThreads* operation assigns each discussion thread in the discussion threads collection to one of the clusters generated by the clustering model. The output is a list of discussion threads and clusters. Each row in the list consists of a discussion thread ID and a cluster ID.
getCentroids	The *getCentroids* operation computes the weights of the terms in the cluster centroid vector for each of the derived clusters. The output is a list of the terms and term weights in the centroid vector of each cluster.
Example usage	When the users of the collaborative workspace log to the discussion forum view, they may want to view the discussion threads that their content meets their interests or needs. Invoking the Cluster Discussion Threads service groups the existing discussion threads into clusters (groups) based on their content. A recommendation service can then compute a similarity score between the user preferences vector and the centroid vector of each discussion threads group and then recommend to the user to read and participate in the discussion threads that belong to the group whose centroid has the highest similarity score with the user preferences.
Comments	The user does not interact with the Cluster Discussion Threads service directly. This is a support service (internal mechanism) where the purpose of this service is to group similar discussion threads found in a discussion forum together into distinct groups. The discussion forums that belong to a particular cluster are relatively similar in content. **Current restrictions**: There are no available discussion threads in the Dicode Discussion Forum view by the Use Case users (Use Case 1 and

	2). Therefore, the current prototypical version of the service is demonstrated using discussions collected from an online technical discussion forum. See Appendix A for the detailed demonstration of the service. **Future Extensions and Enhancements**: Future versions of the service will be applied on the discussion threads made by the users of Use Case 1 and 2 within the discussion forum view of the Dicode multi-view workspace. Future versions will also apply different clustering models based on different clustering algorithms (e.g. K-means, Hierarchical clustering (Abonyi and Feil, 2007)). The different models will be compared and the model that derives the best clusters (based on cluster validity evaluation techniques) will be selected for the final implementation. Feature extraction preprocessing will be performed before generating the clusters using Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) to improve the clustering efficiency and the centroids of the generated clusters.
Conformance classes	Not available.
Implementation rules	Not available.
Implementation status	Prototypical version implemented.
UML model	Not available.

Due to the current unavailability of discussions by Dicode communities in the Discussion-forum view of the Dicode collaboration space, the prototypical version of the service has been demonstrated using discussions collected from an online technical discussion forum. In Appendix A, a detailed demonstration of the service can be found.

2.2.2 Identify Discussion Forum Topics

In Section 7.6 of deliverable D2.2 - The Dicode Approach, the community modelling and user profiling services have been introduced to support Dicode Collaboration and Decision Making Support services. These services will be able to use the output of the community modelling and user profiling services to improve the sense-making and decision making of the user within Dicode communities.

The 'Identify Discussion Forum Topics' service is aimed to enable the user with a Dicode community to make more sense of the discussions that exist in the Discussion-forum view of the Dicode collaboration space by taking a cluster of discussions generated by the "Cluster Discussion Threads" service (described in Section 2.2.1) as an input and produces a list of identified topics with their weights, which the discussions in that cluster are about. The output of the service (topics and topic weights) is expected to help the users of the discussion

forum to make more sense of the existing discussions. By showing the users a "topic cloud" that depicts the identified topics in different font sizes that are proportional to the computed topic weights, the user is expected to identify the main theme of the discussion in that cluster. Therefore, the user will be able to focus his reading and participation in the discussions that belong to the cluster that he identified, based on the identified topics, as the most suitable for his interest.

The following abstract description of the service is given:

Name	Identify Discussion Forum Topics
Standards	
Description	The Identify Discussion Forum Topics service takes a cluster of discussion threads as input from the Cluster Discussion Threads service and identifies the most prominent terms (topics) in the discussion threads that belong to that cluster. The identified topics can be used by the users to identify the main theme the discussion threads in that cluster are about. The Identify Discussion Forum Topics service provides its functionality through the TopicsIdentifier interface. The purpose of this interface is to receive a collection of discussion threads that belong to a particular cluster, perform text pre-processing on the received threads, and identify the topics of these threads.
Interface	TopicsIdentifier
readDiscussions	The *readDiscussions* operation reads the dataset from a storage resource. In the prototypical version of the service, the storage resource is an XML file that contains the textual content of the discussion threads and the cluster ID that each discussion thread is assigned to (See Section B.1 of Appendix B for more details about the input dataset). In subsequent versions, the storage resource will be table / view in a relational database server.
preprocessThreads	The *preprocessThreads* operation receives a collection of discussion threads belonging to a particular cluster and performs text pre-processing on the threads in that collection. The *preprocessThreads* operations include the following text pre-processing Lucene[4] filters: StandardFilter, LowerCaseFilter, and StopFilter (See Section B.3 of Appendix B for more details about these filters). The output is a pre-processed version of the discussion threads
identifyTopics	The *identifyTopics* operation receives the pre-processed version of the discussion threads and identifies the most prominent topics of these discussion threads that belong to a particular cluster. The current version of this operation uses a term frequency – based algorithm to identify the topics (See Section B.2 of Appendix B for a description of the

[4] http://lucene.apache.org/java/docs/index.html

	algorithm). The output is a list of topics and topic weights. Each row in the list consists of a topic that belongs to a cluster and the weight of this topic within that cluster of discussion threads.
Example usage	When the users of the collaborative workspace log to the discussion forum view and focus on the discussion threads that belong to a particular cluster (i.e. the cluster of discussions that is closest in content to their preferences), they may want to summarize these discussion threads or identify the main theme that these discussion threads are about. Invoking the Identify Discussion Forum Topics service identifies the main topics these discussion threads have. A visualization service can take the identified topics and their frequencies and show to the users a "topic cloud" that depicts the identified topics in different font sizes proportional to the frequency of occurrence of these topics. The user will be able to identify how important each identified topic is to the discussion threads based on the font size of that topic, and therefore be able to summarize these discussions and identify the main theme of the discussions. The user may also be able to interact with the identified topics through another visualization service that links each identified topic with a page that shows a list of the discussion threads that contain discussions mainly related to that particular topic. The user can click on the topic link to browse the discussions related to that topic.
Comments	The user does not interact with this service directly. This is a support service (internal mechanism) where the purpose of this service is to identify the most prominent terms (topics) in the discussion threads that belong to a particular cluster of discussion threads. **Current restrictions**: There are no available discussion threads in the Dicode Discussion Forum view by the Use Case users (Use Case 1 and 2). Therefore, the prototypical version of the service is demonstrated using discussions collected from an online technical discussion forum. See Appendix B for the detailed demonstration of the service. **Future Extensions and Enhancements**: Future versions of the service will be applied on the discussion threads made by the users of Use Case 1 and 2 within the discussion forum view of the Dicode collaboration space. Future versions of the service will implement more advanced models to topic identification and topic modeling (e.g. models based on the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003)) than the frequency of occurrence model. The different models will be compared and the model that derives the best topics (based on human evaluation) will be selected for the final implementation.

Conformance classes	Not available.
Implementation rules	Not available.
Implementation status	Prototypical version implemented.
UML model	Not available.

Due to the current unavailability of discussions by Dicode communities in the Discussion-forum view of the Dicode collaboration space, the prototypical version of the service has been demonstrated using discussions collected from an online technical discussion forum. In Appendix B, a detailed demonstration of the service can be found.

2.3 Decision making support services

Dicode's decision making support services is the focus of Task 4.5; they aim towards the formalization of the collaboration to intelligently support stakeholders in decision making activities by enabling the use and exploitation of reasoning mechanisms.

Based on the functional specifications outlined in deliverable D2.2, an initial set of operations supporting collaboration within Dicode have been designed and developed. In particular, the decision making support services implement the formal-view of collaboration workspaces that has been outlined in deliverable D2.2.

2.3.1 Decision making support

Decision making support services aim at providing the necessary operations to implement the formal-view of collaboration workspaces, as outlined in deliverable D2.2. The formal view of a collaboration workspace permits only a limited set of discourse moves for a limited set of message types whose semantics is fixed and system defined. In addition, the formal view of a collaboration workspace can be associated with reasoning algorithms that are able to calculate which proposed solution is currently prevailing or which position has been defeated.

The formal view of collaboration workspaces is the third option in which collaboration workspaces can be operated. The other two are the forum and mind-map view, whose initial version has been reported in deliverable D4.1.1. While the forum and mind-map views of collaboration workspaces aim at supporting collaboration towards sense-making, the formal view aims at supporting collaboration towards decision making.

Since all three abovementioned views constitute an integral part of the collaboration workspace concept, where in addition the transformation between views is possible, the decision making support operations have been implemented within the Collaboration Service, which has been partially presented in deliverable D4.1.1. A different interface (addressing the needs of decision making support) provides the relevant operations when the workspace is operated in formal view. This interface is presented below.

Name	Collaboration Service
Standards	REST (Fielding, 2000)
Description	In addition to what is described in deliverable D4.1.1, the Collaboration Service provides operations that permit users to conduct formal argumentative discourses. Such formal argumentative discourse is provided by the formal view of collaboration workspaces. The service provides operations to allow users to configure the formal view and use them to engage into collaborative activities towards decision making. The Collaboration Service provides functionality for supporting the formal view of collaboration workspaces through the following interface: • *FormalCollaboration Interface:* The purpose of the FormalCollaboration interface is to provide all operations related to the configuration and use of the formal view of workspaces. The FormalCollaboration interface does not provide operations for creating and managing collaboration workspaces spaces, as such tasks are handled through the *Workspace* interface that has been described in deliverable D4.1.1.
Interface	*FormalCollaboration*
postIssue	Allows the creation of issues, which correspond to decisions and goals to be made, open to dispute. The operation takes as input the ID of the workspace where the issue is discussed along with a textual description of the decision to be reached. Upon successful completion, the textual description of the issue appears on the argumentation tree as the root of the collaboration.
postAlternative	Allows the creation of alternatives, which constitute potential solutions, choices or course of actions for the issue being discussed. The operation takes as input the ID of the issue to which the alternative is considered a solution, the title of the alternative that will appear on the workspace, and optional inputs such as comment, URL and attachments that aim at providing more details about the alternative being posted. Upon successful completion of the operation, the alternative appears as a child

	element of the issue which is specified.
postPosition	Allows the posting of a position which is either in favor or against an alternative or other position. The operation takes as input the ID of the alternative or position to which the position responds, the title of the position, an indication whether it supports or is against the alternative and optionally a comment, URL and attachment aiming at providing greater detail of the position being posted. Upon successful completion, the position appears as a child element to the alternative to which it responds.
addPreference	Allows setting up a preference relation, which are used to weight reasons for and against of a certain alternative, in a specific workspace. A preference relation is a tuple of the form *[position, preference_relation, position]*, where the preference relation can be "*more (less) important than*" or "*of equal importance to*". The operation takes as input the two IDs of the positions, along with the preference relation (which can be one of the set {more important than, less important than, of equal importance to}). Upon successful completion, a new preference is created and used by the workspace's reasoning mechanism.
setReasoningEngine	Enables associating a reasoning algorithm with a particular workspace. Within Dicode, a finite set of reasoning algorithms will be developed each of which is identified by a unique name. The operation takes as input the ID of the workspace and the name of the reasoning algorithm. Upon successful completion, the specified algorithm is associated with the workspace and activated only when the workspace is operated in the formal view.
Example usage	A web application will provide the necessary user interface through which the previously mentioned operations can be executed by end users. In general, all Dicode use cases that require support for collaboration towards decision-making will be able to use the above operations. In particular, users of the Dicode use cases can use the operations provided by the workspace interface (presented in deliverable D4.1.1) to create and configure new workspaces where the collaboration will take place. Depending on their needs, they may deploy the workspace either in forum or mind-map view, where they are able to upload and process the available collaboration items via the respective interfaces. If decision making capabilities are required, users may transform a workspace into formal view, through which they may deploy the above operations.
Comments	The initial version of the service reported provides a proof-of-concept implementation of all the presented operations, which nevertheless have bugs and malfunctions. Such issues will be

	addressed in future versions of the services, along with a more thorough evaluation of their usefulness.
Conformance classes	Not available.
Implementation rules	Not available.
Implementation status	Prototypical version implemented.
UML model	Not available.

The **source code** of the initial version of the collaboration service, implementing the above operations, can be found at the following Subversion repository:
https://anivas.cti.gr/websvn/listing.php?repname=ftel&path=/dicode/trunk/src/dll/

In Appendix C, an **example** of how the above operations can be used to cover the decision making support needs of the Dicode use cases is presented.

3 Development Process

The initial version of the previously described services have been developed using the guidelines outlined in deliverable D5.1.1 - Standards and guidelines for development. In particular, the Java Guidelines coding convention has been used to format the authored code. In addition, design patterns have been deployed to solve common software design problems.

To manage code changes, the involved project partners have installed on their site the Subversion code repository, which hosts the code of the initial version of the services they develop.

4 Future Work

This deliverable presents the initial version of the decision making support services that have been designed and developed in the context of Tasks 4.2, 4.3 and 4.5, representing a first implementation of the functional specifications outlined in deliverable D2.2. Future work will concentrate on improving and enriching the above services in order to fully address the needs of the Dicode project. Specifically, with respect to the implementation of the decision making support services, future work will focus on a number of issues which include:

- Correcting bugs and malfunctions of the initial version of the implemented services.
- Assessing the initial version of the developed services against the functional specifications. The purpose of this action is to see how the initial version of the services must be changed in order to properly support the Dicode use cases. This includes identifying which operations must change their functionality as well as which operations must be added in order to fully address the needs of the use cases.

5 References

Abonyi J. and Feil B. (2007). Cluster Analysis for Data Mining and System Identification. Birkhäuser Verlag AG, Berlin.

Bell, R. M., Koren, Y., Volinsky, C. (2007). The BellKor solution to the Netflix Prize, Netflix.com, available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.9009

Blei D., Ng A., Jordan M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning, Research, 3: 993–1022.

Davies, D. L. and Bouldin, D. W. (1979), A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1, 224-227.

Dicode Deliverable D2.2 . (2011). The Dicode Approach User requirements, conceptual integrative architecture, agile methodology and functional specifications.

Dicode Deliverable D3.1.1 . (2011). The Dicode Data Mining Framework (initial version).

Dicode Deliverable D3.2.1 . (2011). The Dicode Data Mining Services (initial version), (Forthcoming).

Dicode Deliverable D4.1.1 . (2011). The Dicode Collaboration Support Services (initial version), (Forthcoming).

Dicode Deliverable D5.1.1 . (2011). Standards and guidelines for development (initial version).

Fielding, R. T. (2000). Architectural styles and the design of network-based software architectures, PhD thesis, University of California, Irvine, Chair-Richard N. Taylor, accessible from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm .

Huttenhower, C., Hibbs, M., Myers, C., Troyanskaya, O.G. (2006) A scalable method for integration and functional analysis of multiple microarray. Bioinformatics, 22(23):2890-2897.

Landauer, T., and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104 (2), 211-240.

Mrowka, R., Lieberneister, W., Holste D. (2003). Does mapping reveal correlation between gene expression and protein-protein interaction? Nat Genetics, 33:15-16.

Pirolli, P. and Card, S. (2005). The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis. In Proceedings of the International Conference on Intelligence Analysis.

Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K. (1993). The Cost Structure of Sensemaking. In Proc. SIGCHI, ACM Press, NY, USA, 269-276.

Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B. (2008). Merging two gene-expression studies via cross-platform normalization. Bioinformatics, 24(9): 1154-1160.

Tusher, V.G., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, in Proc. Natl. Acad. Sci. USA 2001, 98(9): 5116-5121.

W3C – Web Services Architecture. (2004). http://www.w3.org/TR/ws-arch/

Appendix A: Cluster Discussion Threads Service – Demonstrator

This section describes the experiments that have been carried out to demonstrate the prototypical version of the "Cluster Discussion Threads" service, which is part of the Semantic-driven collaboration monitoring mechanism.

Recommendation Support
We aimed to demonstrate the "Cluster Discussion Threads" service by running an experimental study on a set of discussion threads collected from an online technical discussion forum, WebProWorld [5]. The objective is to group similar discussions into distinct clusters. The output of the service is a list of the discussion threads and the clusters that each discussion thread belongs to. The output also produces the cluster centroid vector for each generated cluster, which is a vector that contains the average weights of the terms that exist within the discussions of that cluster. These service outputs (assignment of discussions to clusters, and cluster centroids) can be the foundation for a content-based discussion recommendation service. The discussion recommendation service can compute the similarity score between the user preferences vector and the centroid vector of each cluster. The cluster that its centroid has the highest similarity score with the user preferences vector will be recommended to the user so he may better participate in the discussions of the forum.

Outline
Two pilot experiments have been performed on discussions for demonstration. In the first experiment, the titles of the discussion threads have been used for clustering, whereas in the second experiment, both the titles and the first posts have been clustered. The objective is to check whether including the posts in addition to the titles produce better cluster distribution of discussions. We firstly describe the dataset used as input to the experiments. Secondly, we describe the details of each of the experiments, including data pre-processing, building the clustering model, evaluating the generated clustering models to select the best model for the further steps, and producing the cluster centroids. Finally, we list the URLs of the input and output files that are generated by each service experiment. These URLs link to the output files that can be downloaded from the Dicode Project website.

A.1 Data Set: Online Technical Discussions in WebProWorld

WebProWorld is a large collection of technical discussion forums where each forum has threads that focus on particular IT-related categories. For instance, WebProWorld contains sub-forums that enable registered participants to seek advices and share information about "computer assistance", "search engines", "webmaster, IT and security", and "e-Commerce". WebProWorld provides these sub-forums as a service to its members, to provide a friendly environment, where members can help one another and exchange ideas, tips, news, and information. For example, in the sub-forum "Computer Assistance", users may post questions, answer questions, involve in discussions, seek advices, or share advices about the different problems that IT users face in a daily basis, including hardware, networking, operating systems, web servers, programming, security and others. The thread shown in Figure A.1 for instance was initiated by a user who has a problem with the signal reception

[5] http://www.webproworld.com/webmaster-forum/forum.php

in her house and thus, seeks advice from fellow participants on what the problem might be and how to deal with it.

Figure A.1 An example discussion thread in the "Computer Assistance" sub-forum in WebProWorld

A.1.1 Data Collection

We randomly retrieved a collection of 200 discussion threads from four sub-forums in WebProWorld. These sub-forums are namely: SEO (Search Engine Optimization) forum, IT forum, Computer Assistance forum, and e-Commerce forum. We randomly collected 50 discussion threads from each sub-forum.

A.1.2 Data Attributes

The data attributes for our experiments are:

- Forum ID (1 = SEO, 2 = IT, 3 = Computer Assistance, 4 = e-Commerce forum)
- Discussion thread ID (1 to 200)
- Title of the discussion thread: usually a brief statement that is written by the creator of the thread and represents the main theme of what the thread is going to be about. (e.g. "Hard Disk Problem")
- The creation date of the discussion thread
- The author name of the discussion thread
- The body of the thread, which includes the textual content of the first post written by the creator of the discussion thread

Note: For our experiments, only the first post that is written by the thread creator has been collected in the body because the RSS feeds provided by the forum do not store the reply posts created by the other participants. In the future, we aim to evaluate the service on a complete discussion threads that include the reply posts as well as the titles and original posts. For that, we will consider discussion threads from other technical forums that provide complete RSS feeds, such as the Microsoft Developer Network forums[6] as well as the discussion forums of the Dicode workspace when the communities of use case 1 and 2 start using the Dicode multi-view collaborative workspace.

[6] http://social.msdn.microsoft.com/forums/en-US/categories/

Figure A.2 depicts the data attributes for an example discussion thread parsed from an RSS feed collected from the "Computer Assistance" sub-forum. The figure shows (from top to bottom) the thread ID, the title, the thread creation date, the forum ID, the post ID, the username of the post author, and the body content of the post.

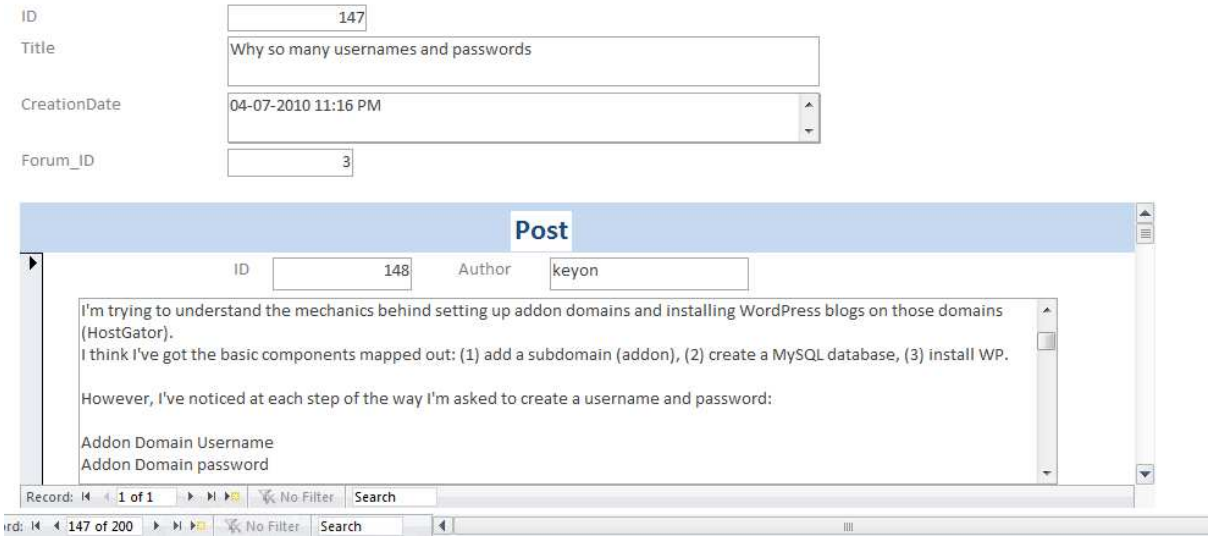

Figure A.2 Data Attributes of an Example Computer Assistance Discussion thread

A.2 Implementation Tools

We have used the following tools for our experiments:

- The Java SE Programming Language for parsing the RSS files that contain the discussion threads and storing them into a database
- MS Access database to store the retrieved and parsed content
- The RapidMiner [7] data mining platform to perform clustering of the collected discussion titles and threads
- MS Excel to store the output of the experiments

A.3 Experiment #1: Clustering Discussion Thread Titles

In the first experiment, we demonstrate the service by grouping the discussion threads into clusters based on the titles of the collected discussion threads. The titles of the discussion threads are retrieved from the database and input as a data source for a clustering model. The service builds and trains a clustering model and assigns each title instance to a particular cluster ID. The output of the service is a list of title IDs and cluster IDs, where each row refers to a particular discussion thread title and the cluster that it belongs to. The service also derives the cluster centroid vectors of all the generated clusters. The output is a list of the top *n* terms and their weights within the centroid vector of each cluster, retrieved in a descending order by their weight (from the term having the maximum weight to the term having the minimum weight).

[7] http://rapid-i.com/content/view/181/190/lang,en/

A.3.1 Data Pre-Processing

We designed a RapidMiner process for data pre-processing, model building, and model evaluation. The following data acquisition and text pre-processing operators have been applied as a data pre-processing phase into the process:

- Read Excel: This operator is used to read the dataset from Excel spread sheet files. For demo purposes, we imported the input titles of the discussion forums from the database into Excel spread sheet files.
- Data to Documents: This operator transforms each discussion thread title within the data set to a document. This step is necessary to treat each title as a separate instance (observation).
- Process Documents: This is the operator that contains the text pre-processing pipeline. It takes the following parameters:
    - Term vector creation weighting scheme. TF/IDF has been chosen as the scheme for weighting each term in the thread title. TF/IDF (Term Frequency / Inverse Document Frequency) is the number of times the term appears in the document multiplied by a function of the inverse of the number of documents in which the term appears.
    - Terms prune method. This parameter specifies if too frequent or too infrequent terms should be ignored for word list building and how the frequencies are specified. The "Absolute" value has been chosen to ignore terms that appear in less than two titles and more than 199 titles.
- The pipeline of the Process Documents operator includes the following sub-operators depicted by figure A.3:

Figure A.3 The Text Pre-processing Pipeline to preprocess discussion title

    - Tokenize. This operator splits the text of a document into a sequence of tokens based on non-letter characters between the terms.
    - Transform terms to lower cases.
    - Filter stop words. Removing insignificant terms such as prepositions and pronouns.
    - Filter tokens by character length. Terms are considered if they have a minimum of two characters.

The resulted output of running the text pre-processing step is a term vector representation for each thread title, as well as a term list showing the frequency of total occurrence for each of the remaining terms (after pruning) and the number of titles the term appeared in. Figure A.4 depicts a snapshot of the term vector representation output, whereas figure A.5 depicts a snapshot of the term frequency of occurrence output.

Row No.	A	access	account	advice	allow	article	awareness	bad	bing	boot	building
1	1	0	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	1	0	0	0	0	0
3	3	0	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0	0
5	5	0	0	0	0	0	0	0	0	0	0.573
6	6	0	0	0	0	0	0	0	0	0	0
7	7	0	0	0	0	0	0	0	0	0	0
8	8	0	0	0	0	0	0	0	0	0	0
9	9	0	0	0	0	0	0	0	0	0	0
10	10	0	0	0	0	0	0	0	0	0	0
11	11	0	0	0	0	0	0	0	0	0	0
12	12	0	0	0	0	0	0	0	0	0	0
13	13	0	0	0	0	0	0	0	0	0	0
14	14	0	0	0	0	0	0	0	0	0	0
15	15	0	0	0	0	0	0	0	0	0	0

Figure A.4 Term Vector Representations of the Thread Titles

Word	Attribute Name	Total Occurences	Document Occurences
ecommerce	ecommerce	13	13
google	google	11	11
security	security	11	10
site	site	10	9
website	website	10	10
links	links	8	8
cart	cart	7	7
help	help	7	7
problem	problem	7	7
seo	seo	7	7
advice	advice	6	6
search	search	6	6
domain	domain	5	5
internet	internet	5	5
password	password	5	5
server	server	5	5
software	software	5	5
tool	tool	5	5
using	using	5	5
account	account	4	4

Figure A.5 Term List showing Frequency of Term Occurrences

A.3.2 Building the Clustering Models

The *k*-Means algorithm has been used to build and train a number of clustering models and generate the clusters of the thread titles for each model. The Euclidean distance measure has been used to determine the title similarity. To determine the best number of clusters, five models have been built, each with a particular number of clusters, ranging from two to six. Figure A.6 depicts the final pipeline for the clustering process.

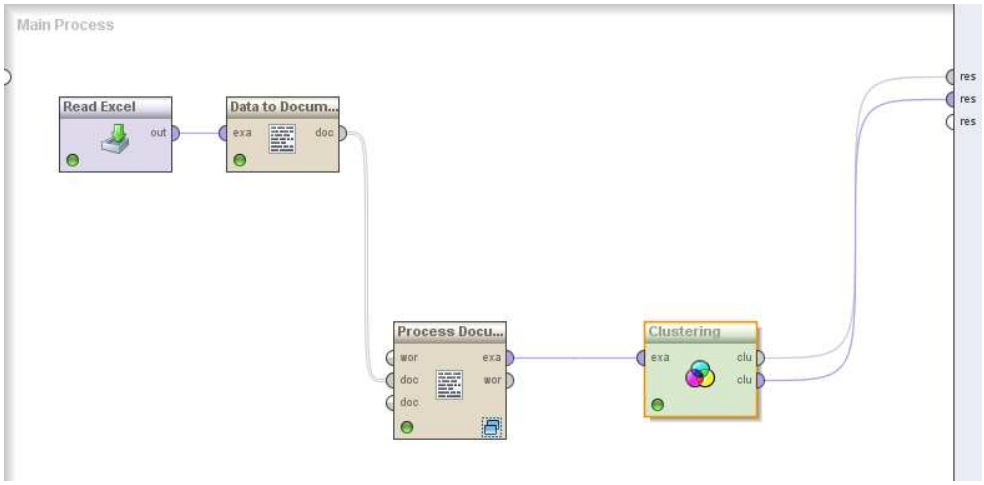

Figure A.6 The operator pipeline for the clustering process

A.3.3 Cluster Evaluation

Two RapidMiner operators have been used to evaluate the trained models in order to select the best one for implementation:

- The 'cluster distance performance' operator: used to validate the performance of each clustering model having *k* clusters based on cluster centroids. The average within cluster distance is calculated by calculating the average of the distances between the cluster centroid vector and all vectors of the titles within the cluster. The model that has the smallest overall centroid-instances distance average is selected. The overall average can also be normalized by dividing the average within each cluster by the number of instances in that cluster and then sum all the resulted averages. The 'cluster distance performance' operator also calculates the Davies–Bouldin index (Davies and Bouldin, 1979) within each generated model. Small values of Davies–Bouldin indices correspond to clusters that are compact, and whose centroids are far away from each other. Consequently, the number of clusters that minimizes Davies–Bouldin index is taken as the optimal number of clusters.
- The 'Item Distribution Performance' operator: evaluates k-means cluster models based on how well the discussion thread titles are distributed over the clusters. This operator calculates its performance in the sum of squares case very simply. The number of titles in each cluster is divided by the total number of titles in all clusters. This is squared and the values for each cluster are summed. For a situation where one cluster dominates and the others clusters are very small in comparison, this value will tend to 1. For the opposite situation, where the clusters have equal numbers of titles, the value tends to 1/N where N is the number of clusters.

Table A.1 lists the generated clustering models and their average centroid-instances distance, normalized average centroid-instances distance, the Davies–Bouldin index, the normalized Davies–Bouldin index, the item distribution rate, and the offset of the distribution rate to normal distribution.

Model ID	No of clusters	Avg. centroid-instances distance	Normalized Avg. centroid-instances distance	Davies–Bouldin index	Normalized Davies–Bouldin index	Item Distribution Rate	Offset to Normal Distribution
1	2	0.792	0.007	2.124	0.019	0.878	0.378
2	3	0.777	0.007	3.153	0.029	0.749	0.416
3	4	0.764	0.007	4.348	0.04	0.511	0.261
4	5	0.755	0.007	4.08	0.037	0.399	0.199
5	6	0.735	0.007	3.622	0.033	0.444	0.277

Table A.1 Cluster Validity Measures for the created clustering models

In table A.1, the best value for each measurement is depicted in yellow. Although the model with six clusters (model #5) gave the smallest average centroid-instance distance value, the distances for all the models are close to each other and become identical when they are normalized. The Davies–Bouldin index and its normalized value both indicate that the inter-cluster distance between the titles is highest when the number of clusters is two. However, the distribution rate for that model shows that there is a large, dominating cluster and a very small cluster, which do not reflect the actual dataset distribution since they have been retrieved from four categories equally (50 threads from each category). The title distributions across the clusters for the fourth model having five clusters show the best (smallest) offset

value to normal distribution. Hence, it is chosen for cluster profiling and topics identification.

A.3.4 Item / Cluster Distribution

For selected model ID 4 (k=5), the title distribution is as follows: (Total no of titles: 200)

- Cluster 0: 20 titles
- Cluster 1: 20 titles
- Cluster 2: 12 titles
- Cluster 3: 119 titles
- Cluster 4: 29 titles

A.3.5 Cluster Centroids

In addition to assigning each discussion title to a cluster, the service also produces the cluster centroid vector for each cluster. The cluster centroid vector is the average of all the points in the cluster for each term. The elements having the highest *n* values in each cluster's centroid are the terms that are mostly occurring within the thread titles of each cluster. Figure A.7 depicts up to 20 identified topics having the highest values in each cluster centroid vector.

The identified centroid terms provide a good representation for the clusters. For example, the list of terms for cluster #0 centroid contains terms that are highly relevant to the e-Commerce theme, such as “ecommerce”, “shopping”, “cart”, “account”, and “payment”. If there is a user X whose user preferences vector includes these terms with relatively higher weights than the other terms in his preferences, the similarity score between the centroid vector of cluster #0 and the user preferences vector of user X will be relatively high in comparison with the similarity scores between his user preferences vector and the centroid vectors of the other clusters. Therefore, user X will be recommended the discussions that belong to cluster #0.

Cluster_0	
Topic	Weight
ecommerce	0.416
site	0.241
redirect	0.072
solution	0.070
shopping	0.066
cart	0.057
time	0.042
bing	0.041
website	0.037
improve	0.035
url	0.035
levels	0.034
design	0.034
problems	0.034
integration	0.031
web	0.031
article	0.030
account	0.029
server	0.028
payment	0.027

Cluster_1	
Topic	Weight
problem	0.287
website	0.255
internet	0.096
chrome	0.096
explorer	0.096
awareness	0.070
google	0.066
security	0.045
hacked	0.042
traffic	0.041
ebay	0.040
entire	0.035
access	0.034
company	0.030
pages	0.030
page	0.029
domain	0.028
engine	0.025
mobile	0.025
optimization	0.025

Cluster_2	
Topic	Weight
advice	0.312
tool	0.153
pst	0.143
search	0.137
outlook	0.133
results	0.104
repair	0.104
store	0.066
software	0.053
testing	0.053
recovery	0.052
file	0.052
password	0.046
engines	0.045
boot	0.041
ecommerce	0.039
google	0.037

Cluster_3	
Topic	Weight
help	0.039
security	0.039
google	0.028
system	0.023
commerce	0.023
free	0.022
using	0.022
hosting	0.021
online	0.020
software	0.020
password	0.019
panda	0.019
pc	0.018
wordpress	0.018
link	0.017
cloaking	0.017
weird	0.017
please	0.016
computer	0.015
question	0.015

Cluster_4	
Topic	Weight
links	0.167
seo	0.167
server	0.079
facebook	0.072
domains	0.072
domain	0.069
bad	0.068
good	0.054
cart	0.052
email	0.047
testing	0.046
name	0.044
account	0.041
shopping	0.040
sites	0.035
increase	0.028
paid	0.028
class	0.028
network	0.027
looking	0.026

Figure A.7 Up to 20 identified topics in each title cluster

A.3.6 Input / Output Files.

- The input file is an Excel worksheet (AppendixA_A.3_Input.xls) that contains a list of the discussion thread titles. Each row contains the title ID and the title content for each of retrieved discussion threads.

The file can be downloaded using the following URL: http://www.dicode-project.eu/sites/default/files/AppendixA_A.3_Input.xls

- The output file is an Excel file (AppendixA_A.3_output.xls) that contains two worksheets:
    - Cluster Assignment: This worksheet contains a list of title IDs and Cluster IDs, where each row represents a thread title and the cluster it is assigned to by the service.
    - Cluster Centroids: This worksheet contains a list of up to 20 terms for each derived title cluster. The terms are listed in descending order based on their weights on the cluster centroid, from the topic having the highest value to the topic having the lowest value.

The file can be downloaded using the following URL:
http://www.dicode-project.eu/sites/default/files/AppendixA_A.3_output.xls

A.4 Experiment #2: Clustering Discussion Threads (Titles and First Posts)

In the second experiment, we demonstrate the service by grouping the discussion threads into clusters based on both the titles and the first posts that are written by the thread creator. The titles and posts represent the discussion thread for every collected discussion and are retrieved from the database and input as a data source for a clustering model. The objective is to see whether the clustering model can produce better thread / cluster distribution when using the discussion posts with the titles than when only using the discussion titles in the clustering process. Similar to experiment #1, the service builds and trains a clustering model and assigns each thread instance to a particular cluster ID. The output of the service is a list of thread IDs and cluster IDs, where each row refers to a particular discussion thread and the cluster that it belongs to. The service also derives the cluster centroid vectors of all the generated clusters.. The output is a list of the top *n* terms and their weights within the centroid vector of each cluster, retrieved in a descending order by their weight (from the term having the maximum weight to the term having the minimum weight).

A.4.1 Data Pre-Processing

We used the same RapidMiner process for data pre-processing. However, in the pipeline of the Process Documents operator, we added the 'Generate *n*-Grams' operator in order to detect multi token terms (n-Gram terms) within the posts of the discussion threads. A term n-Gram is defined as a series of consecutive tokens of length n. The term n-Grams generated by this operator consist of all series of consecutive tokens of length n. We chose n = 2 to detect terms that consist of binary tokens (bigrams) that frequently occur in the discussions. The resulted output of running the data pre-processing operators is a vector representation of each discussion thread, consisting of both single terms and two-gram terms, with the frequency of the total occurrence for each of the remaining terms, after pruning the outlier terms, and the number of discussions each term appeared in. Figure A.8 depicts a snapshot of the output for the term total frequency of occurrence.

Attribute Name	Total Occurences
seen	10
seems	13
seem	5
seeing	7
see	23
security_essentials	5
security	42
secure	5
section	3
seconds	5
search_results	7
search_engines	6
search_engine	9
search	50
se	4
scratch	3
says	5
say	16
sales	5
said	16

Figure A.8 Term Vector Representations of the Thread Discussions (Titles and Posts), with bigram Terms detected

In the figure, some bigram terms that occur frequently in the discussions can be depicted, such as "security essentials", "search results", and "search engines".

A.4.2 Building the Clustering Model

We used the *k*-Means algorithm and the Euclidean distance measure to determine the similarity of the discussions and generate the clusters of discussions. We applied the same approach to determine the model with the best number of clusters, training five models, each with a particular number of clusters, ranging from two to six.

A.4.3 Cluster Evaluation

The 'cluster distance performance' and the 'Item Distribution Performance' operators have been used as in experiment #1 to compute the average within cluster distances, the Davies–Bouldin index, and the item distribution offset values for each of the generated clustering models.

Table A.2 lists the generated clustering models and their average centroid-instances distance, normalized average centroid-instances distance, the Davies–Bouldin index, the normalized Davies–Bouldin index, the item distribution rate, and the offset of the distribution rate to normal distribution.

Model ID	No of clusters	Avg. centroid-instances distance	Normalized Avg. centroid-instances distance	Davies–Bouldin index	Normalized Davies–Bouldin index	Item Distribution Rate	Offset to Normal Distribution
1	2	0.958	0.001	8.092	0.009	0.503	0.003
2	3	0.949	0.001	8.297	0.009	0.4	0.067
3	4	0.941	0.001	7.438	0.008	0.264	0.014
4	5	0.935	0.001	6.964	0.008	0.218	0.018
5	6	0.926	0.001	6.621	0.008	0.195	0.028

Table A.2 Cluster Validity Measures for the created clustering models

A.4.4 Evaluation Analysis

In table A.2, the best value for each measurement is depicted in yellow. The fifth model with 6 clusters has the minimum average centroid-instances distance and Davies–Bouldin index. However, it does have not the best Distribution Rate difference from the normal distribution. The fourth model with five clusters, on the other hand, has very close average centroid-instances distance and Davies–Bouldin index to the fifth model and a better distribution rate closer to the normal distribution for a 5 cluster model. Therefore, the fourth model (k=5) is selected for cluster profiling and topic identification.

A.4.5 Item / Cluster Distribution

For selected model ID 4 (k=5), the title distribution is as follows: (Total no of titles: 200)

- Cluster 0: 36discussions
- Cluster 1: 40 discussions
- Cluster 2: 36 discussions
- Cluster 3: 62 discussions
- Cluster 4: 26 discussions

A.4.6 Cluster Centroids

Similar to experiment #1, the service also produces the cluster centroid vector for each generated cluster. Figure A.9 depicts up to 20 identified terms having the highest values in each cluster centroid vector.

Cluster_0		Cluster_1		Cluster_2		Cluster_3		Cluster_4	
Topic	Weight	Topic	Weight	Topic	Weight	Topic	Weight	Topic	Weight
links	0.062	problem	0.059	password	0.065	site	0.059	search	0.092
blog	0.057	system	0.050	outlook	0.057	ecommerce	0.054	facebook	0.077
panda	0.050	help	0.047	account	0.048	cart	0.052	domain	0.052
seo	0.050	know	0.046	html	0.048	shopping	0.036	results	0.049
blogs	0.047	website	0.044	pst	0.044	shopping_cart	0.034	bing	0.047
link	0.043	chrome	0.040	windows	0.043	looking	0.028	wordpress	0.047
get	0.042	internet	0.037	access	0.041	products	0.027	site	0.042
sites	0.038	com	0.037	security	0.041	server	0.026	seo	0.041
google_panda	0.037	explorer	0.036	file	0.040	page	0.026	month	0.036
building	0.031	google	0.033	server	0.039	com	0.026	com	0.030
find	0.031	page	0.033	domains	0.035	data	0.025	search_results	0.030
anchor	0.031	xp	0.028	user	0.034	know	0.024	google	0.030
good	0.029	pc	0.028	essentials	0.030	prices	0.023	social	0.029
title	0.028	virus	0.026	outlook_pst	0.030	time	0.023	email	0.029
increase	0.028	install	0.026	setup	0.029	website	0.023	going	0.029
text	0.027	part	0.025	accounts	0.027	pages	0.023	mobile	0.029
url	0.026	free	0.025	get	0.027	research	0.023	company	0.029
tell	0.025	google_chrome	0.024	exe	0.026	budget	0.022	history	0.029
pr	0.024	idea	0.024	tool	0.026	customers	0.022	photos	0.028
www	0.024	drive	0.022	pst_file	0.025	want	0.022	content	0.028

Figure A.9 The 20 most prominent identified topics in each discussion cluster

A.4.7 Improvement

There has been a significant improvement in the distribution of discussions across the clusters in all the generated models compared with the title models generated by experiment #1. This is clearly shown in table A.2 by the smaller "offset to normal distribution" values compared to the same measurement in table A.1. The discussions in experiment #2 are almost equally distributed across the generated clusters with no dominating cluster. This aligns with the fact that discussions have been collected in equal sizes from different discussion categories as described in section A.1.1. Including the posts in addition to the titles clearly improved the clustering process.

A.4.7 Input / Output Files.

- The input file is an Excel worksheet (AppendixA_A.4_Input.xls) that contains a list of the discussion titles and posts. Each row contains the title ID, the title content, and the first post for each of retrieved discussion threads.

The file can be downloaded using the following URL: http://www.dicode-project.eu/sites/default/files/AppendixA_A.4_Input.xls

- The output file is an Excel file (AppendixA_A.4_output.xls) that contains two worksheets:
    - Cluster Assignment: This worksheet contains a list of thread IDs and Cluster IDs, where each row represents a discussion thread and the cluster it is assigned to by the service.
    - Cluster Centroids: This worksheet contains a list of up to 20 identified terms for each derived thread cluster. The terms are listed in descending order based on their weights on the cluster centroid, from the topic having the highest value to the topic having the lowest value.

The file can be downloaded using the following URL:
http://www.dicode-project.eu/sites/default/files/AppendixA_A.4_output.xls

Appendix B: Identify Discussion Forum Topics Service – Demonstrator

This section describes the experiments that have been carried out to demonstrate the prototypical version of the "Identify Discussion Forum Topics" service, which is part of the Semantic-driven collaboration monitoring mechanism.

Sensemaking Support
Sensemaking refers to the iterative process of building up a representation of an information space that is useful for achieving the user's goal (Russell et al., 1993). A number of related models have emerged, including the one by (Pirolli and Card, 2005). They propose a model of sensemaking that consists of a series of interconnected loops. The *foraging* loop in this model involves tasks such as searching and filtering information, gradually leading to the identification and organization of relevant knowledge. The "Identify Discussion Forum Topics" service takes a cluster of discussions generated by the "Cluster Discussion Threads" service as an input and produces a list of identified topics with their weights, which the discussions in that cluster are about. The output of the service (topics and topic weights) is expected to help the users of the discussion forum to make more sense of the existing discussions. By showing the users a "topic cloud" that depicts the identified topics in different font sizes proportional to the computed topic weights, the user is expected to identify the main theme of the discussion in that cluster. Helping the users in identifying the right cluster of discussions is considered to be supporting the foraging loop of the sensemaking process.

Outline
Two pilot experiments have been performed on the discussions for demonstration. In the first experiment, clusters of discussion titles have been used as input to identify the topics of discussion in each cluster. In the second experiment, the clusters of discussion titles and first posts have been used to identify the discussion topics. The objective is to examine the identified topics for the two experiments to see the applicability of the service. In this demonstrator, we firstly describe the data sets used, and secondly describe the Hadoop MapReduce model that we used for the experiments. Thirdly, we describe Apache Lucene filters used to pre-process our input data. Fourthly, we describe the experiments giving snapshots and analysis of the identified topics in each experiment. Finally, we list the URLs of the output files containing the discussion topics identified by each service experiment. These URLs link to the files that can be downloaded from the Dicode Project website.

B.1 Data Set: Clusters of Online Technical Discussions in WebProWorld

As described in Appendix A, two experiments have been conducted on an online technical discussion forum, WebProWorld, to demonstrate the "Cluster Discussion Threads" service. The two experiments generated two outputs. The first output is clusters of the forum discussions based on the titles of the discussion threads, whereas the second output is clusters of the same discussions based on both the titles and first posts of the thread creators. For the purpose of demonstrating the "Identify Discussion Forum Topics" service, we use the cluster assignments derived from the "Cluster Discussion Threads" service on the same discussions. Figure B.1 depicts a snapshot of the dataset used.

In this figure, each row in the dataset represents a discussion thread, consisting of the thread ID, title, first post content, the cluster ID based on the thread title (the Title_Cluster column), and the cluster ID based on the thread title and first post (the Thread_Cluster column). The dataset has been retrieved from the database, transformed into an input XML file. In the first experiment, the service takes the discussion titles and titles' clusters as input and generates the topics for the discussions based on the clusters of titles. In the second experiment, the service takes the discussion titles, discussion posts, and threads' clusters as input and generates the topics based on the clusters of the whole discussions.

ID	Title	Content	Title_Cluster	Thread_Cluster
1	Is internal linking going to help my SEO?	Hi Friends,	cluster_4	cluster_4
2	Article Submission	Can anyone give advise on article submission d	cluster_3	cluster_0
3	Organic traffic on website	I want to increase the organic traffic of my web	cluster_1	cluster_0
4	Any Tool Which Can Show Search Volume By Month	If I want to check any keyword or search volu	cluster_2	cluster_4
5	Best web 2.0 sites for building back links	Hi Janeth. I think you may be able to help me. I	cluster_4	cluster_0
6	SEO tutorial	can anyone suggest me a good tutorial link to le	cluster_4	cluster_4
7	The 57 Signals of Personalization	Recently a presentation by Eli Pariser, some ne	cluster_3	cluster_4
8	Paid Links Are hurting me!	As of about 30 minutes ago, a competitor of mi	cluster_4	cluster_0
9	How to improve Google Site Indexing?	I know a sitemap and sitemap submission may l	cluster_0	cluster_3
10	Best way to do a redirect on an existing site	For years a site www.companya.com has been a	cluster_0	cluster_0
11	Benefit from 301 Redirect from Older Domain to Brand New One?	If I have purchased a brand new domain name,	cluster_3	cluster_0
12	PR Question	What is the difference in a PageRank 0/10 and C	cluster_3	cluster_1
13	How big is a ' niche' (keyword selection)	How Big is a niche? . . I used to believe that My	cluster_3	cluster_3
14	Navigation Levels and Crawlers	I heard a while ago that it was important to kee	cluster_3	cluster_2
15	Looking for New Keyword Research Tool	Okay...I was loving Market Samurai as a keywor	cluster_3	cluster_3
16	Using Facebook Notes As Your Blogging Platform For SEO	Hello everyone!	cluster_4	cluster_4
17	A company with no website of their own dominates the first 2 pages of Googl	I did a search a few days ago and was shocked t	cluster_1	cluster_4
18	What happened with twitter?	Any body noticed that now twitter PR is 0, how	cluster_3	cluster_0
19	Cache Date of the Website	While I was searching, I wanted to know one w	cluster_1	cluster_1
20	Dynamic URL's and Link Exchanges	I've recently found another seo that has been o	cluster_3	cluster_0

Figure B.1 Input Dataset for the "Identify Discussion Forum Topics" Service

B.2 Using Apache Hadoop MapReduce for Topic Identification in Discussion Threads

Apache Hadoop MapReduce[8] is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. It is a subproject of the Apache Hadoop[9] project, which develops open-source software for reliable, scalable, and distributed large-scale computing. We designed a MapReduce model for the "Identify Discussion Forum Topics" service. For the purpose of demonstrating the prototypical version of the service, we run the MapReduce job in a Hadoop cluster in pseudo-distributed mode, performing the job of identifying topics from each discussion cluster on a single-node cluster. Under the MapReduce model of the service, the topic identification job has been split into a mapping and reducing phases. Table B.1 illustrates the inputs and outputs for the two MapReduce phases.

	Input	Output
Mapping Phase	$<C_X, DT_Y>$	list($<T_1, \omega>$)
Reducing Phase	$<T_{1A}, list(\omega_A)>$	list($<T_2, W>$)

Table B.1 Inputs and Outputs of the Job Identification Mapping and Reducing Phases

Where:

[8] http://hadoop.apache.org/mapreduce/
[9] http://hadoop.apache.org/

- C_X : the X^{th} cluster, where $X \in \{0, 1, \dots, n\}$, and n is the total number of clusters.
- DT_Y : the Y^{th} discussion thread that belongs to cluster C_X, where $Y \in \{1, 2, \dots, m\}$, and m is the total number of discussion threads in cluster C_X.
- list(<T_1, ω>) : a list of initial topics (T_1) and initial topic weights (ω), where each row in the list consists of a term in the discussion, representing a topic, and its initial weight (1.0).
- T_{1A} : The A^{th} unique topic in the topic list (T_1), where $A \in \{1, 2, \dots, z\}$, and z is the total number of unique topics in the topic list (T_1).
- list(ω_A) : list of all the initial weights that exist for the topic T_{1A}
- list(<T_2, W>): list of the identified unique topics (T_2) and final topic weights (W), where each row in the list consists of an identified unique topic and its final weight.

The final weight W for the A^{th} identified unique topic (T_{2A}) in the final topic list (T_2) is determined by the equation:

$$W_{T_{2A}} = \sum_{i=1}^{k} \omega_i$$

where ω_i is the i^{th} initial weight of the topic T_{2A}, and k is the total number of weights for this topic.

The output of the service is the list (<T_2, W>) for each cluster C_X of discussion threads.

B.3 Combining Apache Lucene with Hadoop for Filtering Topics

Apache Lucene[10] is a high-performance, full-featured text search engine library written entirely in Java. Lucene's class library contains a number of Java classes for text tokenization, filtration, and analysis, which can be used to perform data cleansing, pre-processing and transformation. In the mapping phase of the "Identify Discussion Forum Topics" service, a number of Lucene filtering classes have been used in order to reduce the high dimensionality of the dataset and filter out the terms that do not make a significant contribution to topic identification in discussions. These include the following:

- Standard Filter: Works in conjunction with Lucene Standard Tokenizer class, which removes punctuation and splits words at non-character symbols.
- Lower Case Filter: Normalizes the term text to lowercase.
- Stop Filter: Removes words that appear in a provided stop word list from the token stream.

B.4 Experiment #1: Identified Topics in Thread Titles

Figure B.2 depicts the top 10 prominent topics identified in each of four clusters of discussion thread titles generated by the first experiment of the "Cluster Discussion Threads" service described in Appendix A.

[10] http://lucene.apache.org/java/docs/index.html

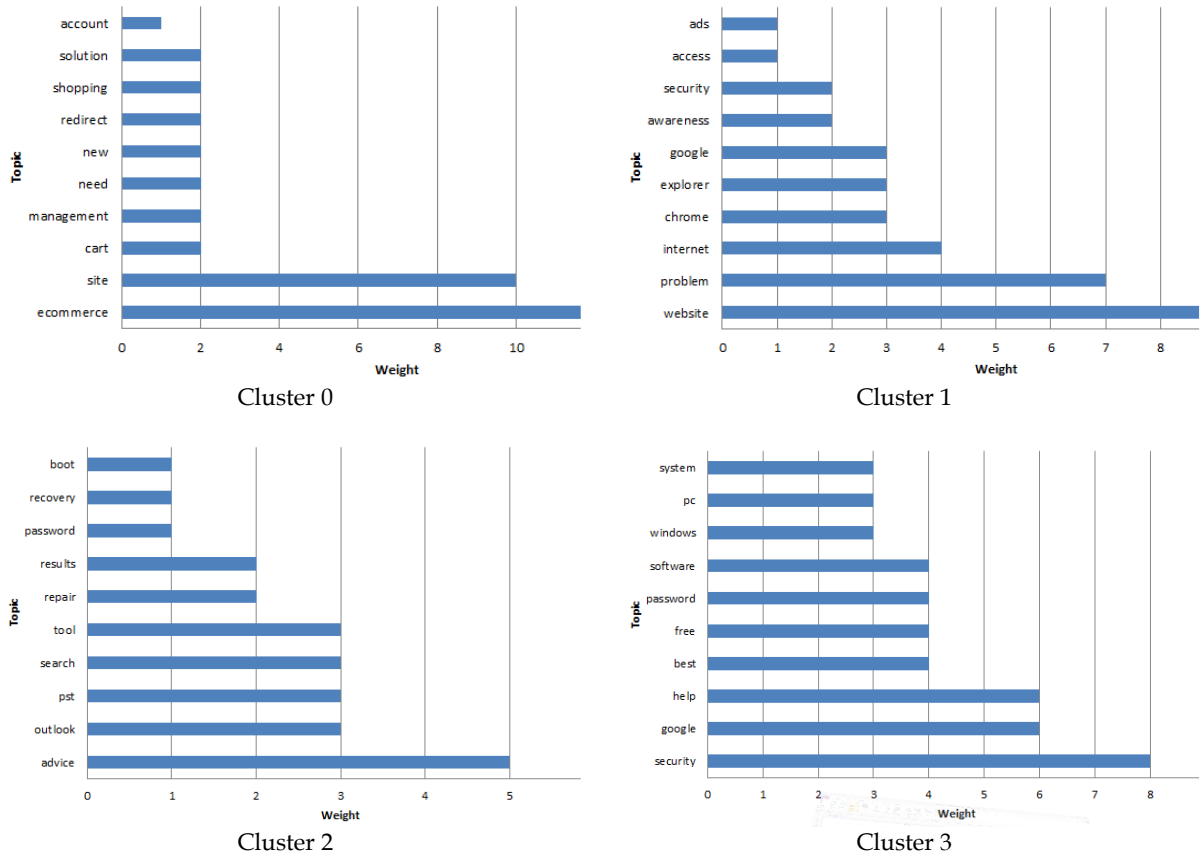

Figure B.2 Top 10 Identified Topics in Four Clusters of Discussion Thread Titles

Analysis

Taking a closer look at the top 10 most prominent identified topics in each cluster, it can be sensed that there is a main theme of discussions for every discussion group. Cluster 0 is mainly about e-Commerce discussion. This can be concluded from the topics that are highly related to the e-Commerce domain, such as "account", "shopping", "management", "cart" and "ecommerce". Cluster 1 contains discussions about Internet security ("access", "security", "awareness", "chrome", "internet", "website"). Cluster 2 discussions particularly address MS Outlook problems ("recovery", "outlook", "pst", "repair", "advice"). Cluster 3 discusses are suitable to those who seek advices related to computers, software and operating systems in general ("system", "pc", "windows", "software", "help", "best"). By looking at a topic cloud visualization for each cluster that shows the identified topics with sizes proportional to their weights in the cluster, a user X who is interested in discussions about e-Commerce can efficiently identify the group of discussions mostly relevant to his information needs (here, cluster 0) and focus on the discussions in that cluster. Similarly, a user Y who seeks advice on an MS Outlook problem can identify cluster 2 as his discussion group of interest when looking at the topic cloud of that cluster.

B.5 Experiment #2: Identified Topics in Thread Titles and Posts

Similarly, figure B.3 depicts the top 20 prominent topics identified four clusters of discussion threads (titles and first posts) generated by the second experiment in Appendix A.

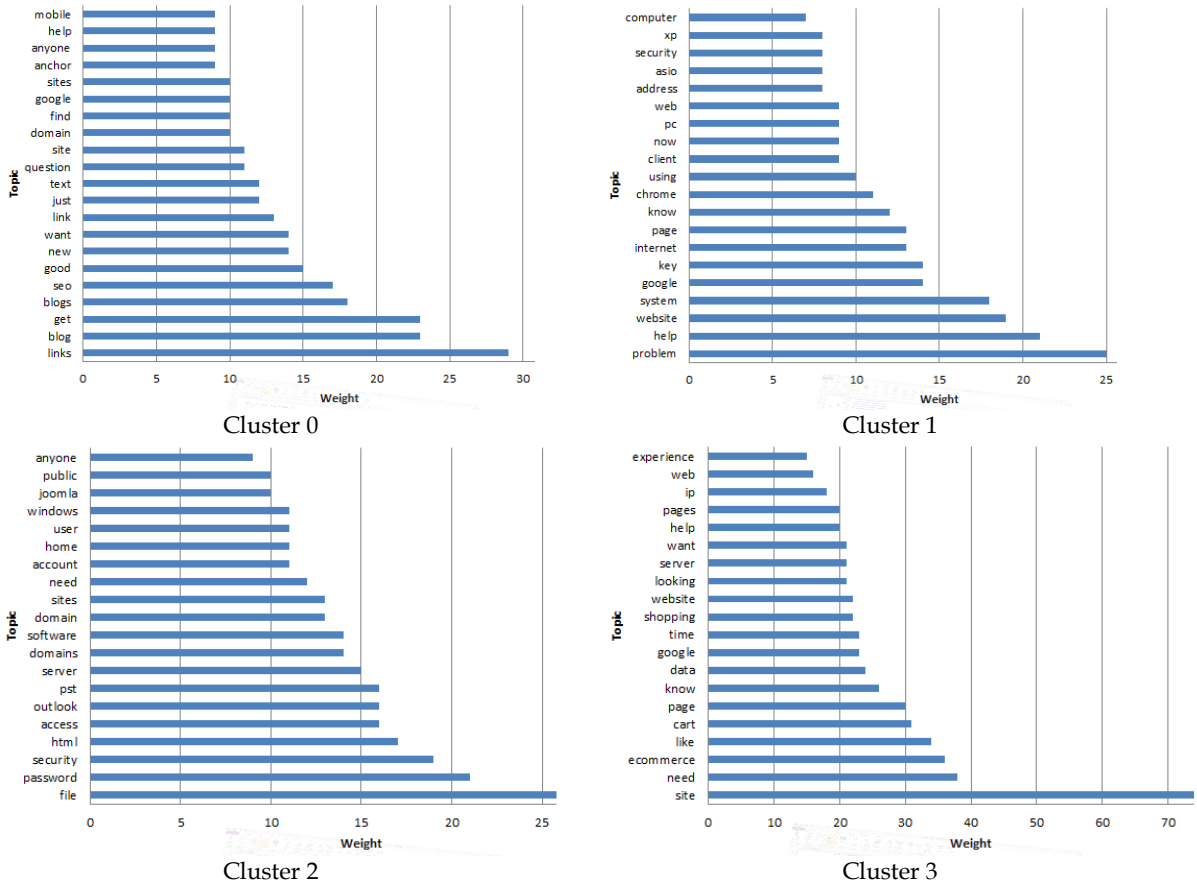

Figure B.3 Top 20 Identified Topics in Four Clusters of Discussion Threads (Title and First Posts)

Analysis

It can be sensed that some clusters contain common terms. For instance, the term "google" exists as an important term in clusters 0, 1, and 3. On the other hand, both clusters 0 and 1 contain general terms that do not provide clear topic distinction between them. Therefore, the current algorithm of the service provides better topic representation for the clusters when using discussion titles to identify the topics of the discussion clusters than when using the whole discussion content (titles and posts). The future implementation of the service will add more advanced algorithms for topic modelling, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and combine these algorithms with the current algorithm for enhanced topic identification from the entire discussions.

B.6 Output Files

The output files for the two experiments are the following Excel files:

- AppendixB_B.4_output.xls: The Excel file containing the output of experiment #1 described in section B.4. This file contains five worksheets: cluster 0 topics, cluster 1 topics, cluster 2 topics, cluster 3 topics, and cluster 4 topics. Each worksheet contains a list of the identified topics for the cluster named by the worksheet, in addition to a bar chart depicting the top 10 most prominent topics for that cluster.
- AppendixB_B.5_output.xls: The Excel file containing the output of experiment #2 described in section B.5. This file contains five worksheets: cluster 0 topics, cluster 1 topics, cluster 2 topics, cluster 3 topics, and cluster 4 topics. Each worksheet contains a list of the identified topics for the cluster named by the worksheet, in addition to a bar chart depicting the top 20 most prominent topics for that cluster.

The files can be downloaded using the following URLs:

- AppendixB_B.4_output.xls:
  http://www.dicode-project.eu/sites/default/files/AppendixB_B.4_output.xls
- AppendixB_B.5_output.xls:
  http://www.dicode-project.eu/sites/default/files/AppendixB_B.5_output.xls

Appendix C: Decision Making Support Operations - Demonstrator

This appendix presents an example of how the service and the related operations described in Section 2.3 are used to achieve the decision making support required in Dicode. In particular, we present and discuss their use in the context of a scenario belonging to Dicode's use case 1, which is relevant to Clinico-genomic research. First, we present a motivational scenario outlining a typical collaboration example, when researchers engage in joint clinic-genomic research work with the aim towards reaching a decision. Although this scenario has also been presented in deliverable D4.1.1, we include its description here as well to aid the reading of this section. Based on this scenario, we then present how the decision making aspects of the scenario are conceived in terms of the decision making support services and operations presented in this deliverable.

C.1 Motivating Scenario

Consider two researchers, Jim and Alice, aiming to investigate which genes or groups of genes are associated with breast cancer disease. Initially, they create a new collaboration session (logbook), where they exchange ideas related to which data sources to use, based on their own data analysis experience and literature knowledge. They search relevant literature via PubMed (http://www.ncbi.nlm.nih.gov/pubmed) using the appropriate search services. Jim has conducted an initial analysis with some in-house gene-expression datasets; however, his findings were not very encouraging, which was attributed to the small sample size (i.e. number of patients) available. He informs Alice about it and suggests potential solutions. The discussion proceeds and finally, in order to overcome the limited sample size problem, they decide to augment their samples with publicly available gene-expression data derived from the GEO (http://www.ncbi.nlm.nih.gov/geo/) and SMD (http://smd.stanford.edu/) databases.

After deciding what data to use, they keep collaborating in order to discuss how the data will be processed. Both suggest solutions, comment on them, and finally decide to use the normalized data for each platform and the UniGene annotation database (http://www.ncbi.nlm.nih.gov/unigene) to uniformly map all genes. Jim knows that there are particular confounding effects in such kind of analysis and for that reason suggests a specific strategy that would account for these effects. Particularly, they decide to first analyze the integrated dataset using the well-known Significant Analysis of Microarrays (SAM) methodology (Tusher et al., 2001). This analysis will serve as a baseline to any further analysis they attempt. Jim is also offering to provide all the necessary R scripts (http://www.r-project.org/) for this initial statistical analysis. In addition, they decide to employ model-based data integration methodologies (Huttenhower et al., 2006) (Shabalin et al. 2008) that have been recently published and claim to perform better than simple data integration techniques (Mrowka et al., 2003).

Some of the models are readily available; however, others need to be coded. Jim offers to write the relevant scripts. Alice, being an experienced programmer, offers to hard code them using parallel programming and various servers available at her department. Parallel and cloud computing will ensure fast results, since they have both agreed that they should apply the selected methodologies to numerous datasets. Their goal is to identify novel or already

reported groups of genes associated with breast cancer disease. In addition, they are interested in comparing the findings of the chosen methodologies to those of the simple analysis conducted by Jim. They decided to quantify and check the statistical and biological significance of their results via the DAVID tool (http://david.abcc.ncifcrf.gov/) and the KEGG database (http://www.genome.jp/kegg/pathway.html).

Both researchers can execute the available services and retrieve the results of the invoked tool (e.g. a scatter plot or heatmap plot). Once the results are available, they engage into interpreting the results in terms of the initial research question. Jim and Alice may need to further elaborate the items considered so far, in order to advance their collaboration towards reaching a final decision. In this case, they need to engage into the task of carefully assessing all available resources in a way that also justifies and proves the final decision.

C.2 Addressing the scenario with the decision making support services

In deliverable D4.1.1 it is shown how in the above scenario the two researchers use the forum and mind-map views of the workspace aiming at supporting sense-making tasks. During these views, Jim and Alice may upload and structure all items on the workspace to assess and increase their utilization.

When aiming towards reaching a final decision, Jim and Alice may exploit additional functionalities of the workspace, which aim at actively assisting them during such decision making process. Such functionalities can be provided by the formal view of a Dicode workspace, which enables the semantic annotation of knowledge items, the formal exploitation of collaboration items patterns, and the deployment of appropriate formal argumentation and reasoning mechanisms.

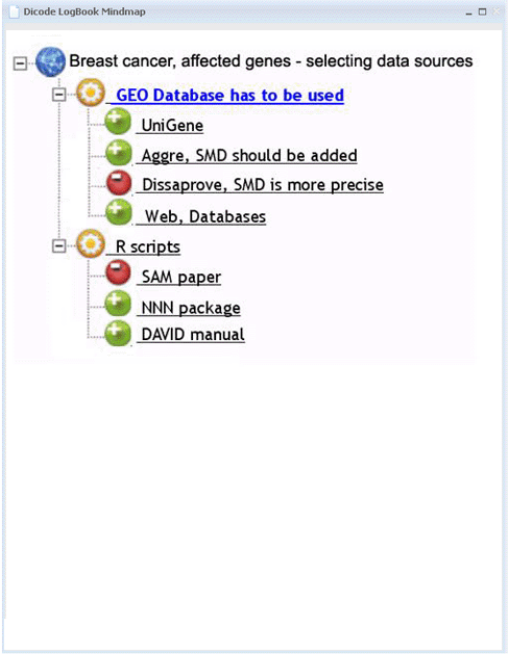

Figure C.1: A formal view of the collaboration workspace

To enable the formal view of a workspace, Jim and Alice, who are working with the mind-map view of the workspace, may call the *tansformWorkspace* operation which transforms the current mind-map into the workspace's formal view (Figure C.1). By switching from mind

map into formal view, existing item types are transformed, filtered out, or kept "as-is" based on a specific set of rues. These rules take also into consideration the item's visual cues.

While a mind map view aids the exploitation of information by Jim and Alice, a formal view aims mainly at the exploitation of information by the machine. This view provides a fixed set of discourse element and relationship types, with predetermined, system interpretable semantics. In particular, issues correspond to problems to be solved, decisions to be made, or goals to be achieved. For each issue, both users may propose alternatives (i.e. solutions to the problem under consideration) that correspond to potential choices. Positions are asserted in order to support the selection of a specific course of action (alternative), or avert the users' interest from it by expressing some objection. A position may also refer to another (previously asserted) position, thus arguing in favour or against it. Jim and Alice may use the *postIssue*, *postAlternative* and *postPosition* operations to continue their collaboration and share on the workspace items of the respective type.

The formal view of workspaces also integrates a reasoning mechanism that determines the status of each discourse entry, the ultimate aim being to keep Jim and Alice aware of the discourse outcome. Both researchers can adjust and configure the reasoning mechanism by calling the *setReasoningEngine* operation to associate the proper (suitable for their needs) reasoning algorithm with the workspace, and the *addPreference* operation to setup preference relations which are taken into consideration by the reasoning algorithm.

Jim and Alice can continue their collaboration in this formal view; each time an element is added to the discussion, this triggers the underlying reasoning mechanism, which informs the team about the most prominent (current) solution. Using the formal view, Jim and Alice receive active support from the system to reach a decision concerning the most appropriate resources for their research.

A more comprehensive presentation of the above scenario in the context of the proposed decision making support service can be found in the following publications:

- Karacapilidis, N., Tzagarakis, M., Christodoulou, S., Tsiliki, G. (2011a). Facilitating Scientific Collaboration in Data-Intensive Biomedical Settings. In: Proceedings of the 10th International Workshop on Biomedical Engineering, Kos Island, Greece, October 5-7 (to appear).
- Ammari, A., Dimitrova, V., Lau, L., Tzagarakis, M., Karacapilidis, N. (2011). Augmented Collaborative Spaces for Collective Sense Making: The Dicode Approach. In: A. Paramythis, L. Lau, S. Demetriadis, M. Tzagarakis, & S. Kleanthous (Eds.), Proceedings of the International Workshop on Adaptive Support for Team Collaboration 2011 (ASTC 2011), Girona, Spain, July 15, CEUR Workshop Proceedings, Vol. 743, pp. 3-13.
- Karacapilidis, N., Tzagarakis, M., Christodoulou, S., Tsiliki, G. (2011b). Facilitating and Augmenting Collaboration in the Biomedical Domain. International Journal of Systems Biology and Biomedical Technologies (IJSBBT) (submitted).