Chapter 3: Classifier Features and Logical Schema

1 Classifier Features

Features needed for the email classifier (Automatic Categorization of emails into folders)	**Features**		
	Email	**Receiver**	**Attachment**
	Email ID [7]	Domain of receiver [1] [9]	Attachment type [1] [9]
	Body Length [1] [9]	Number of CC [1] [9]	Has an attachment [1] [9]
	Content Type [7]	Number of receivers [1] [9]	Number of attachments [1] [9]
	Domain of sender [1] [9]	Number of To [1] [9]	
	Email Date [3] [7]	Receiver Username [1] [9]	
	Email Sender [1] [2] [7] [9]		
	Email Signature [9]		
	Email Subject [1] [2] [9]		
	Is Bcc [1] [2] [9]		
	Is distribution List [1] [9]		
	Language [1] [9]		
	MIME Version [7]		
	Number of punctuation Letters [1] [9]		
	Percentage of capital letters [1] [9]		
	Sender Username [1] [9]		
	Subject Length [1] [9]		
	Wordgram Frequency [1] [2] [9]		

Table 1: Classifier Vs. Features

2 Detailed Classifier Features

Category	Feature	Description	Values	Source (preparation)
Email	Email ID [7]	Identifier for the email message	Long	System maintained primary key
	Domain of sender [1] [9]	Mail Service Provider (gmail.com, hotmail.com, ..etc)	String	Obtained from the sender's email, by taking the substring after the '@' character
	Language [1] [9]	Dominant language in the email body	String	Use a special module to detect the language type of the email body
	Email Sender [1] [2] [7] [9]	Email address of the sender	String	Obtained directly from the email header
	Content Type [7]	Content type	String	Obtained directly from the email header
	Email Date [3] [7]	Date of sending the email represented as the number of milliseconds since January 1, 1970, 00:00:00 GMT	Long	The date is obtained directly from the email header and then transformed to the long representation
	MIME Version [7]	MIME is an internet standard to extend the format of the email to support non-ASCII data	Integer	Obtained directly from the email header
	Bcc [1] [2] [9]	List of email receivers as Bcc	Each recipient is represented as a boolean attribute in the feature tuple.	Obtained directly from the email header
	Number of punctuation Letters [1] [9]	Number of punctuation characters in the body	Integer	Count the number of punctuation letters in the email body

	Is distribution List [1] [9]	Flag to indicate whether the client received this email from a group/distribution list or not	Boolean	Obtained directly from email header
	Email Signature [9]	Signature of the email sender, at the end of the email	String	The signature is extracted from email body
	Wordgram Frequency [1] [2] [9]	Email Wordgram Frequency	Integer	Count the number of wordgrams in the email
	Subject Length [1] [9]	Length of the email subject	Integer	Calculate the size of the subject string
	Percentage of capital letters [1] [9]	Percentage of the capital letters to the letters in the email body	Double	Count the number of capital letters and divide it by the sum of the sizes of all ASCII words in the email body
	Body Length [1] [9]	Size of the email body	Integer	Calculate the size of the body string
	Sender Username [1] [9]	Name of Sender	String	Obtained directly from email header
	Total Number of words	Number of words in the email body	Integer	Calculate the number of words in the email body
	Email Subject [1] [2] [9]	Subject of the email	String	Obtained directly from the email header
Receiver	Receiver Username [1] [9]	Name of receiver	String	Obtained directly from email header
	Number of receivers [1] [9]	Number of email receivers	Integer	Count the number of receivers obtained from email header
	Number of CC [1] [9]	Number of CC recipients	Integer	Count the number of CC recipients obtained from email header

	Number of To [1] [9]	Number of CC	Integer	Count the number of receivers mentioned in the TO header
	Domain of receiver [1] [9]	Mail Service Provider(s) for the receiver(s)	String	Obtained directly from email header
Attachment	Number of attachments [1] [9]	Number of attached files in the email	Integer	Count the number of attachments obtained from the IMAP interface
	Attachment type [1] [9]	Type of attachment	String	The type of the attached file is extracted from attachment informations
	Has an attachment [1] [9]	Flag to denote whether the email has an attachment	Boolean	If the number of attachment is zero, return false, else return true

3 Logical Schema for Classifiers Features: Data Model – ERD and ETS

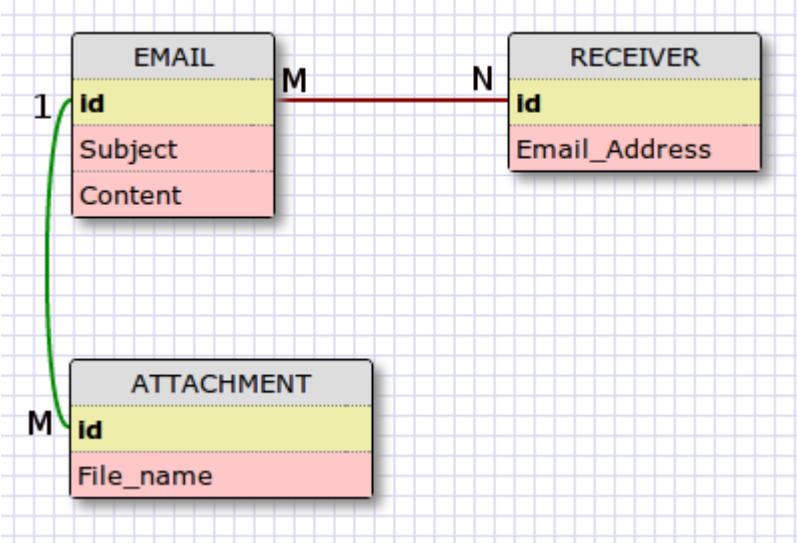

Illustration 1: Logical Conceptual View of the classifier features Data Model (DM)

Chapter 3: Classifier Features and Logical Schema

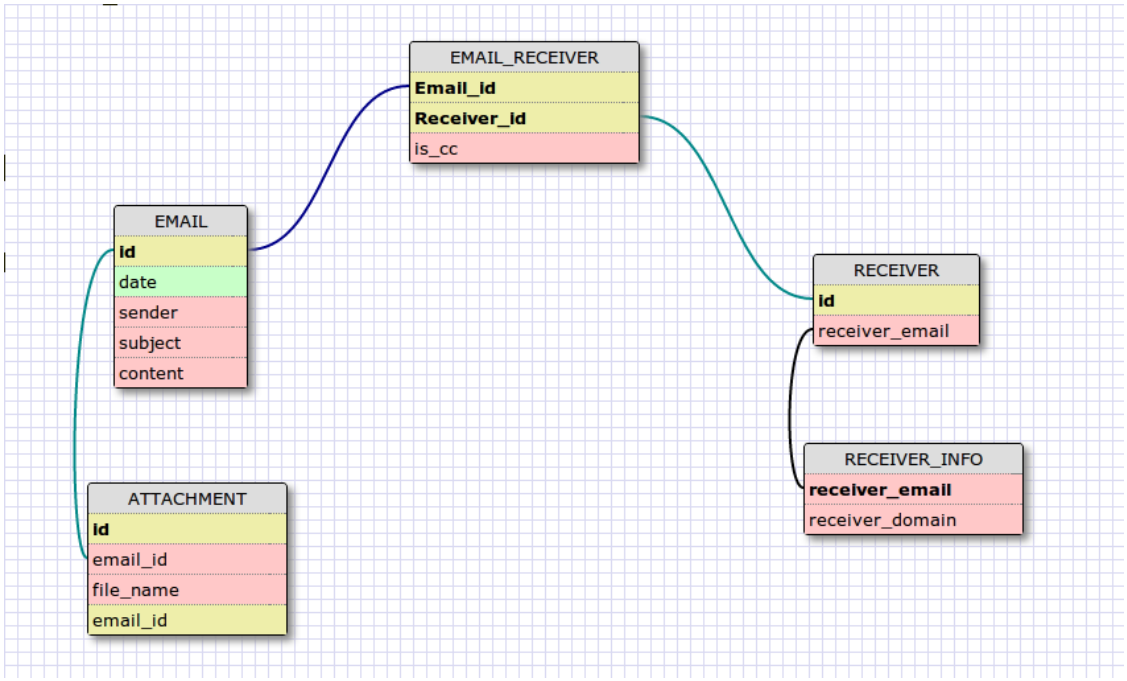

Illustration 2: Mapping DM into Entity Relationship Diagram (ERD)

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 1/1	
Entity: Email	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
Id	Integer	4 Bytes	Primary Key, system maintained, used to identify different emails
date	Date	-	the date the email was received in
sender	String	40 Characters	The email address of the email sender
subject	String	40 Characters	The Subject of the email
content	Text	-	The body of the email

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 1/1	
Entity: ATTACHMENT	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
Id	Integer	4 Bytes	Primary Key, system maintained, used to identify different attachments
email_id	Integer	4 Bytes	Foreign key to EMAIL.id
file_name	String	40 Characters	Name of the attachment

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 1/1	
Entity: RECEIVER	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
Id	Integer	4 Bytes	Primary Key, system maintained, used to identify different recipients
receiver_email	String	40 Characters	Email of the receiver
receiver_name	String	40 Characters	Name of the receiver
receiver_domain	String	40 Characters	Domain of the receiver

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 1/1	
Entity: EMAIL_RECEIVER	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
email_id	Integer	4 Bytes	Composite Primary Key, system maintained, used to identify different email/receiver tuple
receiver_id	Integer	4 Byte	
is_cc	Boolean	1 Byte	Used to indicate whether this receiver is mentioned in cc header or not

4 Tuple of the classifier Features Warehouse (Classifier Input)

email_id
date
sender_email
sender_username
Domain of sender
is_bbc
subject
Subject_length
content
content_mime_version
body_length
signature
number_of_receivers
Percentage_of_capital_letters
Number_of_punctuation_letters
language
has_attachments
number_of_attachments

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 1/2	
Entity: Feature_tuple	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
Id	Integer	4 Bytes	Primary Key, system maintained, used to identify different feature_tuples
email_id	Integer	4 Bytes	Foreign key to EMAIL.Id
date	Date	-	Date of sending the email in the long representation
sender_email	String	40 Characters	Email of the sender
sender_user_name	String	40 Characters	Name of the sender as it appears in the contacts list
domain_of_sender	String	40 Characters	Email Service Provider for the sender
is_bcc	Boolean	1 Byte	Indicates whether the user received this email as a recipient or as a Bcc recipient
subject	String	40 Characters	The Subject of the email
subject_length	Integer	4 Bytes	Length of the email subject
content	Text	-	Email body
content_type	String	40 Characters	Content type of the email
content_mime_version	Integer	4 Bytes	MIME version of the email
body_length	Integer	4 Bytes	Length of the email body
signature	String	512 Characters	Sender's signature at the end of the email message
number_of_attachments	Integer	4 Bytes	Number of attachments in the email

Chapter 3: Classifier Features and Logical Schema

Project: Smart Email	**Subject:** Classifier Features	**Page:** 2/2	
Entity: Feature_tuple	**Date:** Thursday, March 1, 2012	**Analyst:**	
Attribute	**Type**	**Size**	**Validation / desc**
number_of_receivers	Integer	4 Bytes	Number of email recipients
percentage_of_capital_letters	Double	8 Bytes	Percentage of capital letters in the email body
number_of_punctuation	Integer	4 Bytes	Number of punctuation letters in the email body
language	String	40 Characters	Dominant language of the email body
has_attachments	Boolean	1 Byte	True if the email has attachments, false otherwise

5 References

[1] JM Carmona-Cejudo, M Baena-Garcia, Feature extraction for multi-label learning in the domain of email classification, 2011
[2] R Bekkerman, Automatic categorization of email into folders, 2004
[3] S Kiritchenko, S Matwin, Email classification with temporal features, 2004
[4] G Forman, An extensive empirical study of feature selection metrics for text classification, 2003
[5] YF Yi, CH Li, Email Classification Using Semantic Feature Space, 2008
[6] Y Yang, A comparative study on feature selection in text categorization, 1997
[7] Anatomy of an email message, http://www.rickconner.net/spamweb/anatomy.html
[8] SQL designer http://code.google.com/p/wwwsqldesigner/
[9] JM Carmona-Cejudo, M Baena-García, Gnusmail: Open framework for on-line email classification