



BlueSpace Mail Search

Contents

1.1 Index

1.1	INDEX.....	2
1.2	LIST OF FIGURES.....	2
1.3	LIST OF TABLES.....	3
2	ABSTRACT.....	4
3	INTRODUCTION.....	5
3.1	EMAIL.....	5
3.2	EMAIL SEARCH.....	5
3.3	BLUESPACE.....	5
3.4	USER IMPACT.....	5
4	BLUESPACE TECHNOLOGY.....	6
4.1	BLUESPACE RELATIONAL ARCHITECTURE.....	6
4.2	TRANSMAIL AND SUBMESSAGE PATHS.....	6
4.3	BLUESPACE EXTERNAL TRANSMAIL.....	8
4.4	BLUESPACE GOVERNANCE.....	9
5	BLUESPACE TRANSMAIL SEARCH.....	11
5.1	INTRODUCTION.....	11
5.2	MAIL DATA.....	11
5.2.1	Mail data inputs.....	11
5.2.2	Mail data relevance.....	12
5.2.3	Mail data collaborative relevance.....	13
5.3	MAIL ACTIONS.....	13
5.3.1	Mail actions inputs.....	14
5.3.2	Mail actions relevance.....	14
5.3.3	Mail actions collaborative relevance.....	15
5.4	MAIL ACTIONS IMPORTANCE.....	16
5.5	USER ACTIONS.....	17
5.6	SUMMARY.....	17
6	EMAIL COMPARISON.....	18
7	USER INTERFACE.....	19
8	ALGORITHMS.....	28
9	CONCLUSION.....	33

1.2 List of Figures

Figure 1 – Email versus TransMail.....	6
Figure 2 – BlueSpace TransMail scenario.....	7
Figure 3 – Branching TransMail paths.....	8
Figure 4 – TransMail to external contacts.....	8
Figure 5 – Complete compliance model.....	10

1.3 *List of Tables*

Error! No table of figures entries found.

2 Abstract

E-mail was invented in the early 1970s as part of an academic experiment.¹ Whilst its fundamental protocols (such as Simple Mail Transfer Protocol – SMTP) have changed little since their inception in 1982,² real-world usage of e-mail has changed dramatically. According to META Group, 75% of corporate knowledge will reside in email by 2005,³ and 80% of corporate employees prefer email to the phone. Today, business runs on email – it gets used for the vast majority of interactions between people.

Email has catalysed an increase in the speed of business and the volume of information presented to employees. Whilst an average employee five years ago might have received 30 letters and memos each week, many employees today receive 200 emails each day. PWC in the US has 25,000 employees, and recently indicated to BlueSpace that it handles 5 million emails each day.

Many of these emails are non-critical to their recipients, but may be copied to them for 'information'. Email facilitates circulation to many recipients (and thus dysfunctional behaviour) due to its ease of selecting many recipients. The challenge for employees is to navigate through this 'information torrent'. To discover what is actually relevant to them through a sea of noise. This problem is particularly acute for management employees.

BlueSpace employs a relational architecture to add structure to mail content. Each piece of content is audited and managed through its lifecycle in the mail system, and every significant user action is audited. This provides a powerful set of raw data that can be used in relevance algorithms.

Most email systems today have basic search facilities in which each email is given equal weighting. BlueSpace allows searching by keyword, process code and other such criteria. But in addition, BlueSpace uses a multi-dimensional relevance system, that rates the importance of each piece of content based on the following criteria:

1. **User actions** – what did the user do with the mail? If they moved it to a folder or forwarded it then it is likely to be of significant importance.
2. **User collaborative actions** – what did other users do with the mail? Again, their actions indicate their views on the content's importance.
3. **Administrative settings** – a 'Relevance Administrator' can add weightings reflecting the enterprise's view of what is important.

The BlueSpace relevance system does not only rate content – it also rates the importance of each user and other entity in the system. A user can interact with the relevance results by changing a 'slider' (graphical element) on a per mail or per sender basis, to help the system learn.

BlueSpace relevance provides much needed innovation in the user experience of electronic mail. A user can display a folder, but view only content of say 80% relevance and above, so that they can handle the most important mail content first. The search facilities are far in advance of any comparable system. A user can change the relevance algorithm in real-time to improve search results – e.g. "show me content that was read by more people". A user can also view what happened to mail content they have written during its lifecycle, including recipients, read times and the path the content took.

Email has become both the largest unstructured information store and the most comprehensive repository of knowledge in modern corporations. Enterprises and their employees need more effective mechanisms for mining the information in their mail systems. BlueSpace mail search provides such a mechanism. The granularity of the BlueSpace relational mail architecture allows access to the raw data required to deploy a highly relevant search system. BlueSpace allows people to find the content they're looking for in mail.

3 Introduction

3.1 Email

Email was invented in the early 1970s as part of an academic experiment.⁴ Whilst its fundamental protocols (such as Simple Mail Transfer Protocol – SMTP) have changed little since their inception in 1982,⁵ real-world usage of email has changed dramatically. Email is now a mission-critical business tool, and one of the most resource-hungry services provided by IT. Analyst firm Gartner has indicated that the volume of email is growing by 50% every year,⁶ causing companies' bandwidth requirements to double every six months.⁷

SMTP treats each email as a string of text similar to a postcard or telegram that must be passed between computers until it reaches its desired recipients. When a user replies to or forwards on an email, they photocopy all the content in that email and put it into a new email text string with their own comments. This has made email systems into a 'content dumpster', with no structure or ability to audit content, or to mine the information inside the email system.

According to META Group, 75% of corporate knowledge will reside in email by 2005.⁸ But there is no way to mine this data effectively. Most corporations allow users to view only the mail content they have received directly, whilst traditional content management systems have access rights layered over the entire content store. This again reflects the history of email as a person-to-person communications system, rather than a system designed for knowledge management.

3.2 Email Search

Email search is a disparate market, with different products offering varying value propositions. Some companies offer a query-like store as a way to remove content from an active mail server to help such servers scale. Others have attempted to branch into email search from content management systems, but have faced challenges due to the lack of structure in the mail system and the lack of defined access rights. Which corporation would just allow everyone access to all emails? It would be carnage.

3.3 BlueSpace

BlueSpace uses its unique relational technology to provide greater granularity into content in a mail system. By coupling this to its governance model, BlueSpace can determine which pieces of a content a user should be allowed to search. The innovation described in this white paper involves adding advanced analytics into a search system in BlueSpace, causing a dramatic improvement in relevance for the end user attempting to find content.

3.4 User Impact

Much of the innovation in previously published BlueSpace white papers has focused on either enterprise governance and control, or infrastructure efficiency or automation. An obvious question to ask is, "Will people use, adopt and 'fall in love with' the BlueSpace application?" This white paper outlines a powerful addition to the user experience – greatly improved and collaborative search capabilities that help the user navigate their mail. Email may have become the enterprises' core knowledge management repository. But it is ironic that the current, mainstream search facilities for end users in email applications such as Microsoft Outlook are so antiquated and ineffective.

4 BlueSpace Technology

4.1 BlueSpace Relational Architecture

BlueSpace has developed a messaging platform that is very different from existing messaging and mail systems. As introduced in Section 3.1, email photocopies content with each generation of reply and forward, as shown in Figure 1.

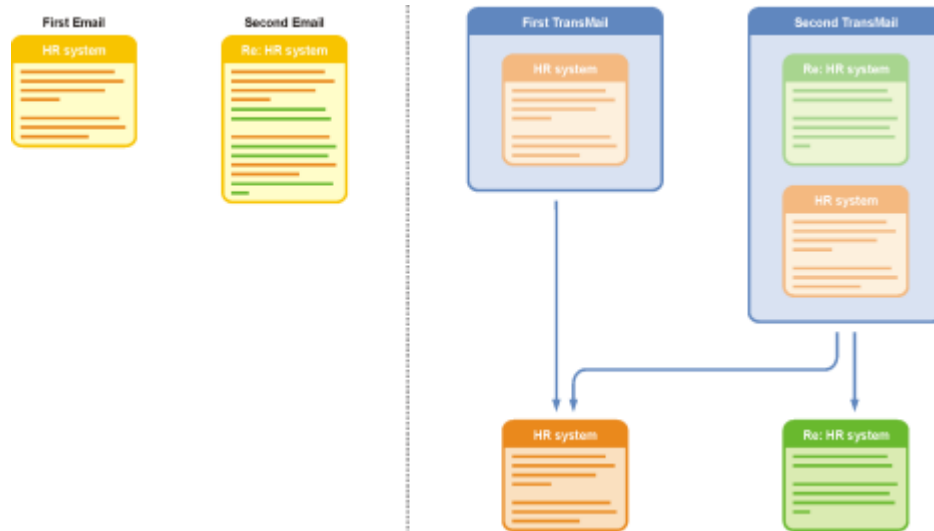


Figure 1 – Email versus TransMail

In the second email, the red message is duplicated without any reference to its origin – there is no link between the first email and the second email. In the BlueSpace TransMail system, the first TransMail points at the red message, and the second TransMail points at both the red and green messages.⁹ Both red and green messages are termed 'submessages' in the BlueSpace environment.

This relational pointing has two powerful outcomes:

1. **Control:** It is far easier to manage and audit the content, enforce access rights and guarantee retention policies are rigorously applied, as there is just a single copy of the message to manage.
2. **Cost saving:** The TransMail system uses far less infrastructure than regular email systems, as it does not photocopy the same piece of content in to many separate places on the infrastructure.

4.2 TransMail And Submessage Paths

An important concept in the world of BlueSpace is that of paths – the way information is moved around a BlueSpace system. Due to the relational architecture of the BlueSpace TransMail application, BlueSpace is able to track what happens to content during its lifecycle, as well as enforcing rules over what is allowed to happen to it.

This section is included to assist the reader in understanding the relational model of BlueSpace, and its importance in search relevance. The user is not required or expected to go to this level of detail when using the application – it's irrelevant to them, they just send mail.

Figure 2 illustrates an example of a series of BlueSpace TransMail communications in which the user Justin Marston is involved.

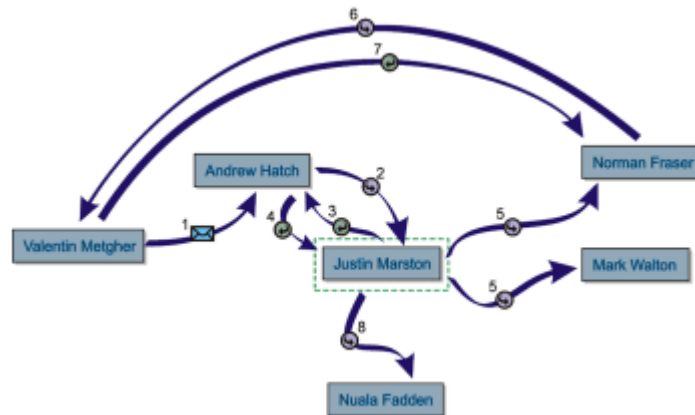


Figure 2 – BlueSpace TransMail scenario

The steps 1 to 8 in Figure 2 are as follows:

1. Valentin Metgher sends a TransMail to Andrew Hatch that contains submessage A.
2. Andrew Hatch forwards the TransMail to Justin Marston, creating a new TransMail that contains B|A (a new submessage B, with history portion submessage A).
3. Justin Marston replies to Andrew Hatch, creating a TransMail that contains C|BA.
4. Andrew Hatch replies to Justin Marston, creating a TransMail that contains D|CBA.
5. Justin Marston forwards the TransMail generated in step 4 to Norman Fraser and Mark Walton, creating a TransMail that contains E|DCBA.
6. Norman Fraser seeks to clarify the situation with Valentin Metgher, sending a TransMail containing F|EDCBA.
7. Valentin Metgher replies to Norman Fraser, creating a TransMail that contains G|FEDCBA.
8. Some time later, Justin Marston forwards the TransMail he received from Andrew Hatch to Nuala Fadden, creating a TransMail containing H|DCBA.

There are three types of submessage with respect to any given TransMail or other submessage:

1. **Current submessage:** The submessage at the top of the chain in any given TransMail is termed the 'current submessage', which will have been written by the user who created the TransMail. In Figure 1, the red submessage is the current submessage in the first TransMail, and the green submessage is the current submessage in the second TransMail. In step 5 of Figure 2, submessage E is the current submessage.
2. **History submessage:** The previous submessages in the chain of any given TransMail (e.g. the red submessage in the second TransMail in Figure 1) are termed 'history submessages'. History submessages are content that was created before the current submessage and its respective TransMail. In step 5 of Figure 2, submessages D, C, B and A are the history submessages.
3. **Future submessage:** For any given submessage, submessages that appear in later TransMails that reference the original submessage are termed 'history submessages'. Thus in Figure 1, the green submessage is a future submessage with respect to the red

submessage. In step 5 of Figure 2, submessages F and G are the future submessages, but submessage H, as the TransMail generated in step 8 does not include submessage E.

Note that a submessage can only have one history path, as it can be the 'current submessage' in only one TransMail. A submessage can, however, have many future paths, as it may be included in many branching forward and reply structures. This is illustrated in Figure 3.

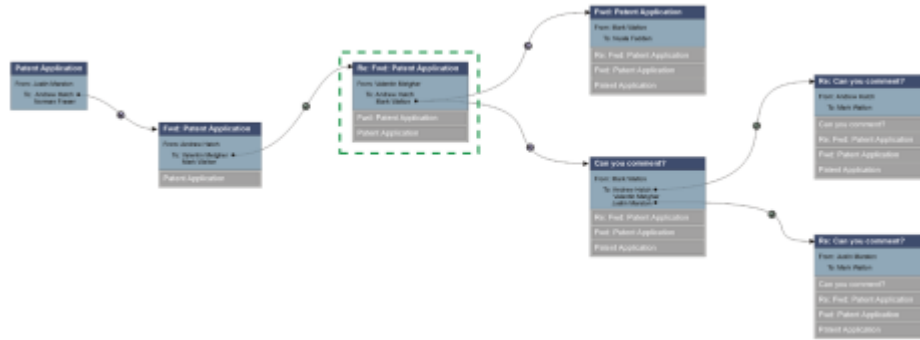


Figure 3 – Branching TransMail paths

All definitions of history and future are thus relative to a given point in time. In Figure 3, the ringed TransMail from Valentin Metgher to Andrew Hatch and Mark Walton is being considered as current. History submessages were written as part of the TransMails to the left, Future submessages were written as part of the TransMails to the right.

Note that a user may have already read content that constitutes both history and future submessages as part of other TransMails when reading a given TransMail.

4.3 BlueSpace External TransMail

BlueSpace can allow TransMail users to send content to external contacts, as shown in Figure 4.

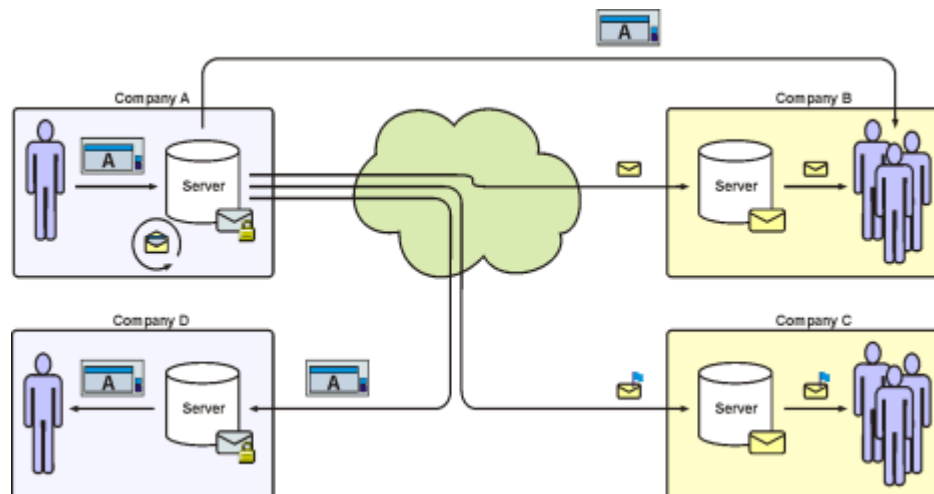


Figure 4 – TransMail to external contacts

Figure 4 illustrates three different methods:

1. A user in Company A sends a TransMail to users in Company B. The BlueSpace server in Company A sends an email notification (with a URL link to the TransMail) to the recipients in Company B. The

recipients in Company B then click the URL to read the TransMail through a web browser.

2. A user in Company A sends a TransMail to users in Company C. The BlueSpace server in Company A wraps the TransMail up in an SMTP envelope, and sends it like a regular email to the email server in Company C. The recipients in Company C view it as normal.
3. A user in Company A sends a TransMail to a user in Company D. The BlueSpace server in Company A realises that the mail server in Company D is a BlueSpace server using BlueSpace Discovery Service.¹⁰ The BlueSpace server in Company A sends the TransMail to that in Company D using a protocol based on web services, and the recipient in Company D then reads the TransMail.

The second scenario, in which BlueSpace wraps a TransMail in an SMTP envelope, is the most complex to manage at the application level. Sending out the email is easy enough, but interpreting any replies can be a challenge, especially in corporations with BlueSpace systems deployed alongside regular email systems. This has been discussed in detail in the BlueSpace Email Structuring white paper.¹¹

4.4 *BlueSpace Governance*

BlueSpace manages content during its entire lifecycle according to rules laid down in the BlueSpace Integrated Governance Management¹² model. This determines how long content is kept for, who can access it, where the data should reside, and the other allowable actions to be carried out by users.

A UML representation of the complete BlueSpace compliance model is presented in Figure 5.

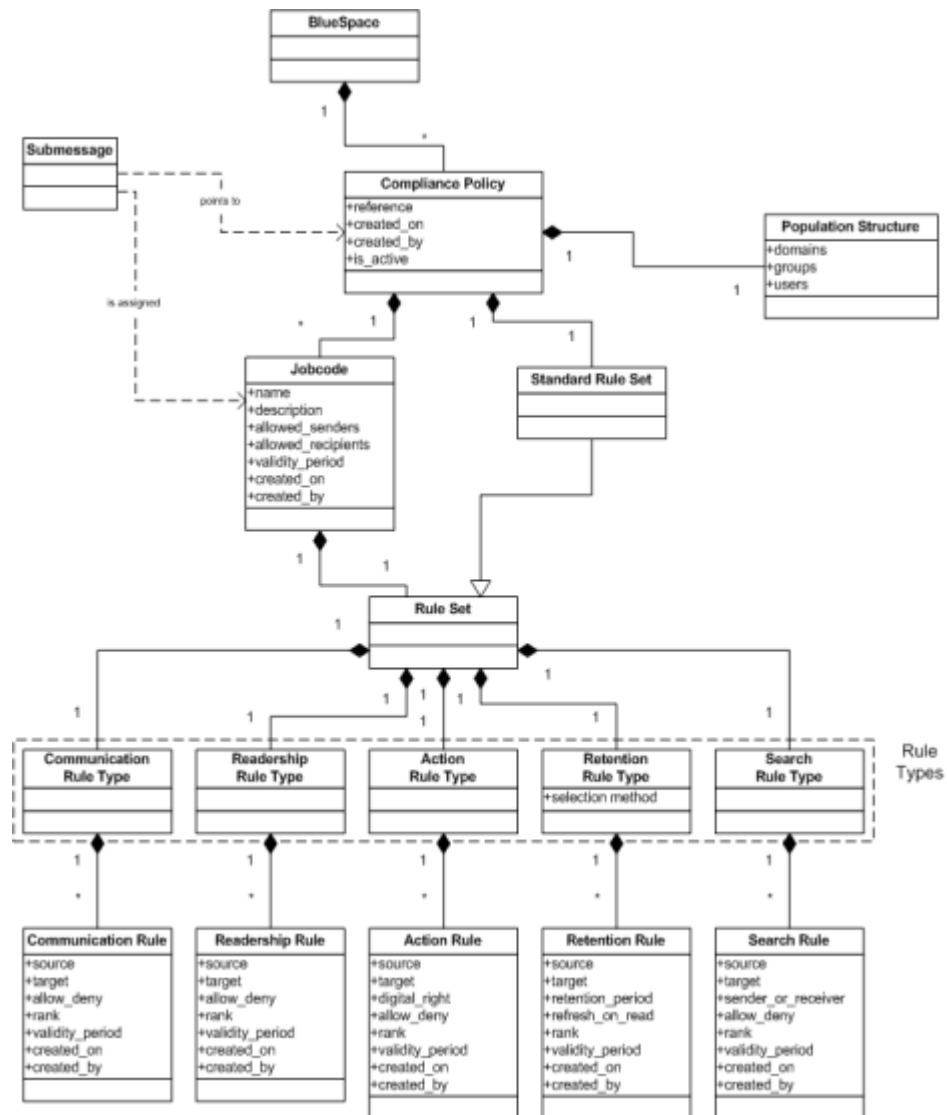


Figure 5 – Complete compliance model

The key aspect for this white paper is the ability of BlueSpace to layer search rules over a content system, determining who has the right to search different pieces of content in the system. For example, in an investment bank, an analyst might have the right to search all content related to analyst processes written by other analysts.

The BlueSpace search rules technology is fully described in the BlueSpace Integrated Governance Management¹³ model.

5 BlueSpace TransMail Search

5.1 Introduction

As described in Section 4.1, BlueSpace separates each TransMail into its constituent message objects, written by different authors. This provides far greater granularity than any other messaging system. The granularity and ease of access to the information is the key foundation supporting the powerful analytics that BlueSpace can apply to a TransMail system. Behavioural information can be easily gleaned from the mail environment and used to help improve content search relevance.

There are two types of information regarding mail content that shall be discussed:

1. **Mail data:** Data in the mail itself, including keywords, creation dates, recipient information, Jobcodes and other meta-type data.
2. **Mail actions:** Information regarding users' actions on mail content, including the number of times it has been read, the number of times it has been forwarded, etc.

In addition, end users can themselves rate various aspects and entities in the system to improve relevance, and can view the entire lifecycle of mail content they have written.

Users interact with relevance in three different ways:

1. New mail – when a user receives a new TransMail, relevance can indicate how important that content is likely to be.
2. Reading mail – when a user is viewing their folder listing or a specific TransMail in the mail view, relevance can indicate how important each TransMail and submessage is likely to be.
3. Searching mail – when a user is searching through TransMail to find something, relevance can allow results to be ranked according to likely interest.

An 'importance' factor is determined for each entity (users, groups, Jobcodes etc) in the system, acting to weight various inputs to the relevance algorithms. Importance is based on historical learning in the system, based on previous actions by users.

5.2 Mail Data

This section is divided into three

1. Mail data inputs – both conventional and BlueSpace-specific criteria that can be used to search content.
2. Mail data relevance – scenarios illustrating how this data can affect search relevance.
3. Mail data collaborative relevance – scenarios involving other pieces of content in determining the relevance of a given piece of content.

5.2.1 Mail data inputs

Many of the criteria for searching mail data in BlueSpace are the same as those in traditional mail search environments:

- Keywords in subject
- Keywords in message body

- Keywords in other fields
- Keywords in attachments
- The date the content was created
- The recipients of the content
- Whether it was sent To:/Cc:/Bcc:
- The sender of the content
- The body size
- The attachment size
- The attachment type

Other criteria are unique to BlueSpace, and leverage the relational architecture of the BlueSpace TransMail system coupled to the strong emphasis BlueSpace places on business process integration:

- Jobcodes – a mechanism for linking an entire submessage (as defined in Section 4.1) to a specific business process.
- Context – a mechanism for linking any text in a submessage to a specific process, either on a user or enterprise-wide basis, as discussed in the BlueSpace Context¹⁴ white paper.
- Groups – sets of users across the system.
- Domains – an X.400-style naming directory across the system.
- Priority – an urgency tag linked to notification processes.
- Sensitivity – a confidentiality tag linked to readership and search rights.
- Validity date – the period of time for which the content is viewable by the user.
- Retention date – the period of time for which the content is retained by the corporation.
- Other Meta-style information stored relationally with respect to the mail.
- Other business processes that are triggered by the mail event.
- Flags for follow up – the sender indicates the TransMail requires action by the recipient.

5.2.2 Mail data relevance

Examples of ways in which these parameters might affect relevance to a given user as part of a search include:

- If the keyword they are looking for is in the body, then it's likely to be relevant to them.
- If a user is looking for a Microsoft Word file, then TransMails with word attachments are relevant to the search.
- If a user is looking for a piece of content they received some time ago from the HR department, which was pretty important, then all of those criteria (date period, priority and the HR group/domain) will be relevant to the search.
- If a user is looking for all content relating to a particular M&A transaction, and that content has been tagged with Jobcodes, then content with the appropriate Jobcode number will be relevant to the search.

- If a user received a TransMail as one of fifty other Cc: recipients, it is less likely to be of relevance to them than if they received the TransMail as the sole recipient.
- If a user is looking for all TransMails that will trigger a financial purchase over the next 30 days in a BlueSpace system deployed for ERP, then the user can search for TransMails that will cause a purchase over the defined date range.
- If a mail is flagged for follow up by the sender, it is more likely to be of significant relevance to the user.

5.2.3 Mail data collaborative relevance

Other scenarios leverage the relational structure of BlueSpace to determine the 'communication context' of the TransMail. These scenarios rely on both both history submessages (older submessages in the chain of the TransMail), and future submessages (submessages written later as part of other TransMails that contain the submessage), as discussed in Section 4.2.

- If a TransMail is received by a user as a reply to a TransMail they themselves sent, then a submessage in the history will map to the 'current submessage' (the one at the top of the chain) in a TransMail in their sent folder. In this case, the incoming TransMail is likely to be of significant relevance.
- If a particular submessage appears many times in a user's inbox, then it is likely to be of relevance to the user.
- If a new TransMail has identical content to a previous one that a user has received, but there is a new current submessage at the top saying something like "did you get this?", then it is likely to be of less relevance to the user.
- If historical submessages in a new TransMail were found to have high relevance ratings, then it is likely that the new submessage will have significant relevance.
- If a submessage A does not contain a given keyword, but many submessages that were sent in replying to and forwarding the submessage A do contain the keyword, then submessage A is likely to still be relevant in a search for the given keyword.
- The longer a given TransMail remains unread, the less likely it is to be of relevance.

Note that these cases do imply user actions, and thus overlap with those given in Section 5.3.

5.3 Mail Actions

This section has been divided into the following subsection:

1. Mail actions inputs – the user actions recorded by the system.
2. Mail actions relevance – scenarios illustrating relevance based on the user's own actions.
3. Mail actions collaborative relevance – scenarios involving other users' actions to determine content relevance.

5.3.1 Mail actions inputs

BlueSpace audits all significant user actions inside a BlueSpace system. These actions are important as they reflect the users' opinions of the content, and the value they place on it. Example of such actions are:

- A user composes a TransMail.
- A user reads a TransMail.
- A user forwards a TransMail.
- A user replies to a TransMail.
- A user deletes a TransMail.
- A user files a TransMail in a custom folder.
- A user prints a TransMail.
- A user copies a TransMail to their clipboard.
- A user opens an attachment.
- A user forwards an attachment.

5.3.2 Mail actions relevance

Mail actions relevance relies on the user's actions to help determine content's relevance for content in all the user's mail folders – both sent and received. Consider the following user action scenarios, in which a user's own actions can be used as an input to a relevance algorithm:

- If a user receives a TransMail from another user, and immediately forwards it to three other users with comments, then it is likely that the content had relevance to them.
- If a user receives a TransMail and replies to it, then the content of the original TransMail is likely to be of significant relevance.
- If a user receives a TransMail from another user, and immediately deletes it, it is unlikely that they perceived the content to be of significant relevance.
- If a user reads a TransMail and then moves it to a custom folder out of the inbox, then it is likely that the content had relevance to them.
- If a user has read a given TransMail many times, then it is likely to be of significant relevance to them.
- The longer the user has a TransMail on screen for, the more likely it is to be relevant to them.
- If a user copies text from a given TransMail to their clipboard in Windows, assuming they have the rights to do so, then this will likely be more relevant than a TransMail in which they did not do so.
- If a user always deletes content from a specific user quickly, then it is unlikely that any content from this user is of significant relevance to them.
- If a user always reads content from another specific user before reading content from other users in any given new set of new mail received into their inbox, then it is likely that content from this specific user is of significant relevance.
- If a user typically only reads short TransMails rather than long ones, then the chances are that short ones are more relevant to that user than longer ones.

- If a user never reads messages sent to a particular group or domain (that they're a member of), then any content sent to that group or domain is likely to be of less relevance to the user.
- If a user never reads TransMails tagged with a particular Jobcode, then this content is likely to be of little relevance to them.
- If a user rarely reads TransMails designated as lowest priority (or reads them last when browsing their inbox),
- If a user sends a TransMail to a lot of users, it is likely to be more relevant than a TransMail they sent to a single user.
- If a user sends a TransMail, and then later forwards it to other users, it is likely to be of more relevance than a TransMail that was sent once but never resent.
- If a user reads a TransMail from their sent folder, then the content is likely to be of significant relevance.
- If a user saves an attachment to their local hard disk or network system, it is likely to be of significant relevance.
- If a user prints an attachment or TransMail, it is likely to be of significant relevance.
- If a user sends a TransMail to their boss, it is likely to be more important/relevant than a TransMail to a junior secretary.
- If a user sends a TransMail with extensive formatting, it is more likely to be important/relevant to them than if they write a few lines of plain text.
- If a user has not accessed the folder in which the content is stored for over a year, then it is likely that the content is not particularly relevant.

5.3.3 Mail actions collaborative relevance

Mail actions collaborative relevance relies on the actions of other users to determine the relevance of content in a given user's folders – both sent and received. In the following scenarios, other users' actions can be used to help determine the relevance of content to a given user:

- If a user sends a TransMail to someone, and they read it, it is more likely to be relevant than if they do not.
- If a user sends a TransMail to a group of users, and most reply to it, forward it or copy it to custom folders, it is more likely to be relevant than if they all delete it immediately.
- If most other users delete a given piece of mail content as soon as it arrives in their inbox, then the user in question will probably not find the content particularly relevant.
- If most other users copy a given TransMail to a custom folder once they have read it, then the user in question will probably find the content to be of significant relevance.
- If a given user is always copying another user on their mail content (e.g. an employee copying their boss) then the content is less likely to be of relevance to them.
- If a user sends a submessage to another user, but the recipient doesn't read it, then this may have an affect on the relevance of the content.

- If the vast majority of the users in a given department read a piece of content sent globally, then the content is likely to be relevant to another user in that department.
- If the majority of other users tend to read content linked to a particular Jobcode, then content attached to that Jobcode is more likely to be relevant to the user in question.
- If the majority of other users tend to read content linked to a particular context, then content attached to that context is more likely to be relevant to the user in question.
- If most users read content from a particular author (e.g. the CEO), then content from this author is typically of greater relevance than content from a different author that no-one ever reads.
- If most users reply to content from a particular author (e.g. the CEO), then content from this author is typically of greater relevance than content from a different author that no-one ever replies to.

5.4 Mail actions importance

In addition to considering content and users' actions on it, the application can also analyse the importance of particular actors and entities when determining their relative weighting in calculating relevance. Such entities include:

- Users
- Groups
- Domains
- Folders
- Jobcodes
- Contexts
- Other business process information
- Any back-end/automated actors (e.g. external computer systems)

In normal business operations, every employee has an importance rating for other employees. For example, it is likely that an employee would find other individuals who report to them and those to whom they themselves report to be of greater importance than individuals working in a different department.

Importance in BlueSpace is also a relative value, respective to each user in the system. For example, users, groups, domains and other entities all have different relative importance values describing a particular Jobcode. Ultimately it is possible to calculate an overall importance value for any given entity in the system, which describes that entity's importance to the entire enterprise.

Importance for any given contact can in fact have two different values. The first value reflects the user's interest in content they received from the contact, and the second value reflects the user's interest in content they sent to the contact. Whilst it might be expected that these values would be very similar, there are scenarios in which they would not be similar. If the head of HR is extremely 'chatty', he/she may have low importance when sending content to other users, but those other users may view the head of HR as being very important when they themselves are doing the communicating.

Varying weightings for relevance inputs with importance means that if a user such as a boss deletes a TransMail, then this will have a greater affect than if a

junior secretary deletes it. Thus different users' actions have different affects on relevance, due to varying importance factors.

Importance has multiple feedback affects, as if user B receives a TransMail from user A, and then user B forwards it to user C who is of high importance (e.g. the CEO), then this affects the importance of all three users.

In an organisation with, say, 100,000 employees, it is likely that any given user will communicate with new contacts each week. Importance is a relative factor for each pair of users' (or entities') interactions, and is stored with respect to each user, as described above. However, when a new contact communicates with a user for the first time, the system can still make a 'best guess' about that new contact's importance based on collaborative importance. The value then changes as the user interacts with this new contact. This first 'best guess' value is important, however, as in a system with a large population, the usability would be hampered if every new contact started 'at par'.

In certain scenarios, an organisation may decide to tie importance to roles rather than specific users. This is particularly likely in highly regimented environments, such as the military, where the role is a greater determining factor than the actual human personality involved. The system is able to make special allowances when a user changes job, as their relevance to others will change quickly, as would be expected.

5.5 User Interaction

In addition to the actions that users perform in the BlueSpace system, BlueSpace can take direct user input to help determine relevance. For example, a user can have a 'relevance slider' for each user in the TransMail system. For example, a user might make their boss 'very relevant' but their junior, chatty sales clerk 'rarely relevant' – these weightings can then be added into the relevance algorithm.

Each piece of content (submessage) can be presented to the user along with a slider that indicates its current relevance score. The user can move this slider, and apply the change in relevance to their own view of that piece of content, or add that change back in to how relevance is calculated across the whole users' mailbox.

In addition to the end user being able to affect relevance directly for their own inbox, certain 'super-user' roles have the ability affect relevance across the entire system. For example, a 'Relevance Administrator' can provide a boost to communications from the HR department. The reason for this is that relevance is not a purely user-centric phenomenon – a separate question relates to relevance for the entire enterprise. Most people may ignore interaction with the HR department, but that doesn't mean that the executive board regard communications from the HR department to be irrelevant.

5.6 Summary

BlueSpace allows users to base their searches on conventional inputs, but adds in new criteria that they can choose to use if they wish. The application also ranks search results according to a relevance algorithm, rather than affording each result equal weighting. In calculating the relevance value, the application also takes into account the importance of each user in the system.

6 Email Comparison

BlueSpace TransMail has a relational mail architecture, in which each piece of content written by a user is managed as a single instance object during its life in the mail system, as described in Section 4.1. The granularity of this architecture allows the application to record all actions by users in the system, ensuring that the application has audited the entire lifecycle of the content.

Regular email does not have this relational architecture – each time a user replies to an email they photocopy the history into a new text string, thus making it very difficult to audit the lifecycle of the mail content.

The BlueSpace application has the ability to analyse regular email content and convert it into a relational system, similar to BlueSpace. This was first introduced in Section 4.3, and has been discussed in detail in the BlueSpace Email Structuring white paper.¹⁵

This approach is less elegant than structuring mail at the front end of the process as in BlueSpace TransMail. When the BlueSpace application structures information once it has already been sent, the criteria for search are those such as:

- Keywords and other standard inputs
- User mail flow actions – compose, reply and forward
- Collaborative user mail flow actions – compose, reply and forward

When the BlueSpace application is integrated with the email environment at the application level (as opposed to being a purely 'downstream' process), then the following criteria are also available:

- User mail management actions – read, delete and file to custom folder.
- User attachment management actions – view, save to disk.

By employing the relational architecture for regular email, BlueSpace allows powerful searching of regular email, with a similar relevance system to that of the BlueSpace TransMail search engine. BlueSpace also layers the access rights policy model over the search results, as described fully in the BlueSpace Integrated Governance Management¹⁶ model.

BlueSpace email search may be used as a standalone knowledge management system, for regular email systems with no TransMail penetration. It may also be used for searching legacy data once a mail system has been transferred to TransMail, or on external email content received through an SMTP gateway in a BlueSpace system with external email capabilities.

7 User Interface

7.1 Introduction

As first presented in Section 3.4, BlueSpace relevance and importance in mail search cause a significant improvement in the end user experience of a mail environment. BlueSpace mail search is semi-automated, in that users actions determine the baseline behaviour of the system, but any given user can interact directly with the inputs into the relevance algorithm to better tune it to their requirements.

This section is dedicated to describing the user interface for BlueSpace mail search, and more specifically, the BlueSpace TransMail search system when delivered through the default web interface.

7.2 BlueSpace TransMail View

There are various core pages of the BlueSpace TransMail web application that include relevance and importance aspects:

- Mail view – when viewing a single TransMail.
- Folder view – a folder listing showing many TransMails.
- Address book – a list of contacts, groups and domains.
- Preferences – a list of information

In addition, the administrative interface presents these statistics to end users.

7.2.1 Mail view

In the TransMail view, a user is looking at a single TransMail, and able to see

its entire content, as shown in Figure 6.

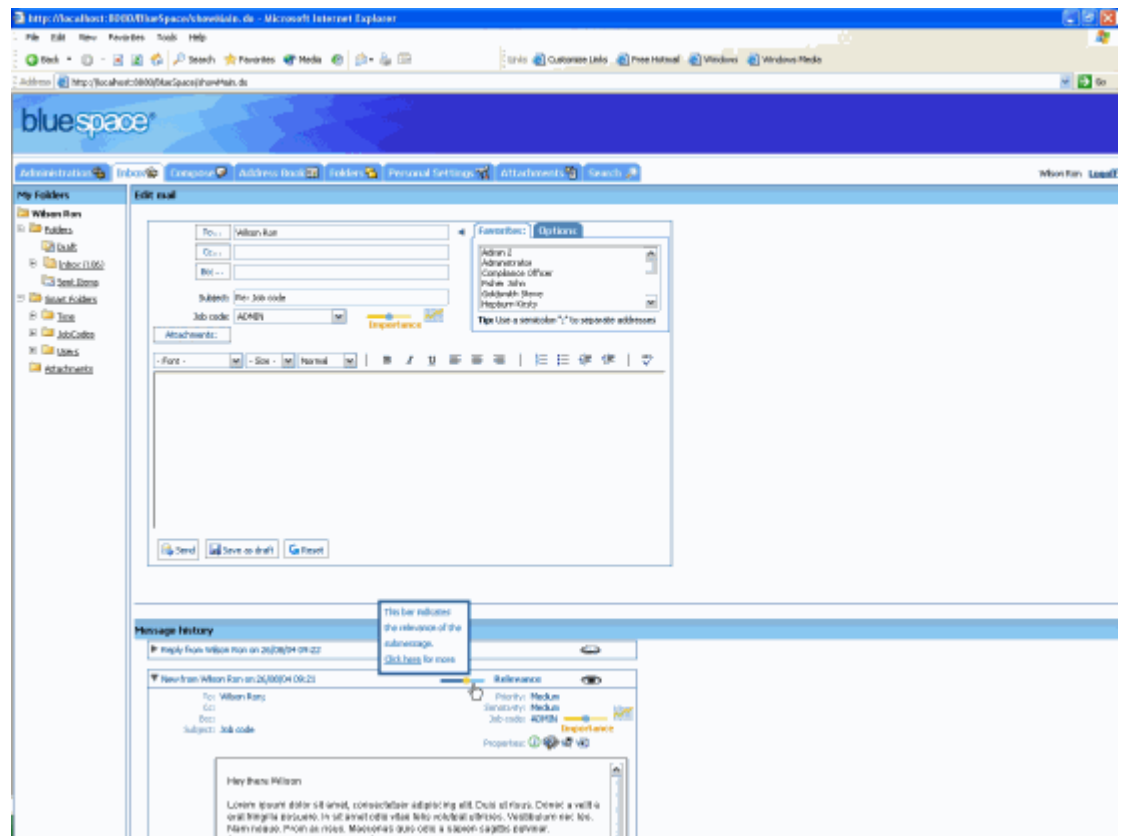


Figure 6 – BlueSpace TransMail view

There are various relevance slides in Figure 6 – one for each submessage in the TransMail view, and one for the entire TransMail, taking into account the current submessage and all history submessages. The user can, if they wish, change these sliders to help the system learn. There is also an icon that presents a popup, indicating why a particular relevance value is as it is – helping the user understand how the algorithm came to its conclusion.

The interface indicates to the user which submessages they have already seen before (e.g. in the case of a reply history), and which are new (e.g. in the case of a forward history). The interface also has buttons next to each submessage in a TransMail that allow the user to see a custom search folder showing every TransMail that contains the given submessage. Another button allows the user to view the lifecycle of any given submessage, visualising its path through the system and all users that have carried out actions on it. Another button allows the user to 'find all content like this' – either at the TransMail or submessage level.

In addition to relevance feedback, the TransMail view can present information on importance. Jobcode importance and other custom created variables that are relatively sparse are given importance sliders as shown in Figure 6. A user can alter the importance slider of a Jobcode, and either lock it at the new value or allow it to recalculate over time. When it comes to users' actions and users'

importance (as well as that of groups and domains), there's a lot of data. Rather than clutter the interface with all this information by default, the application allows the user to click the details icon that then displays a new box that presents this information. This is illustrated in Figure 7. It shows the sender and each recipient of the TransMail, their actions on the content (forwarding the TransMail, replying etc) and their importance.

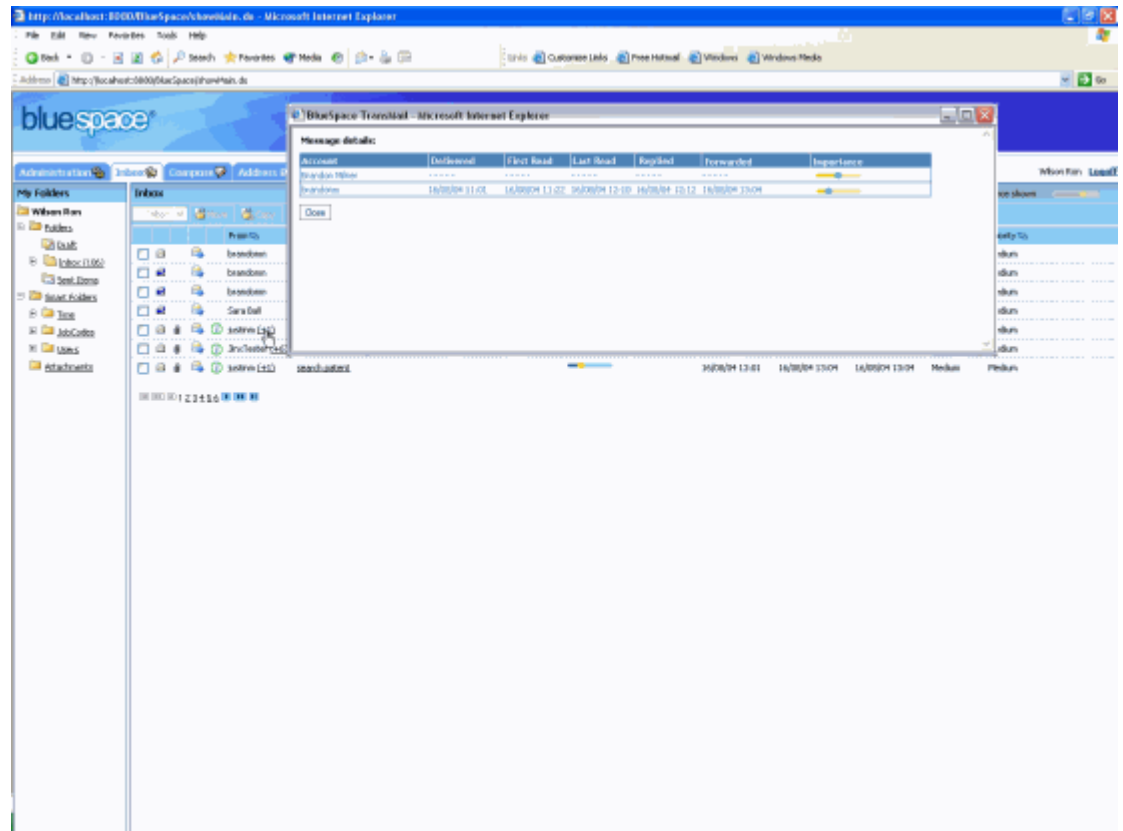


Figure 7 – BlueSpace TransMail user view

Importance is again presented through sliders that the user can change – to tell the application that their boss is important. When a user makes a change, they can either lock the value with respect to the system (so it the user stays at that relative importance) or they can allow it to find its equilibrium value. The latter is like stressing a spring in a spring lattice – over time it will return to its lowest energy state, though this is likely to be different to the initial state as the system is continually changing over time.

Importance clearly changes over time – for example, when a user is interacting strongly with another user on a given project, then that other user will be of greater importance at that point than a year after the time when the project has finished. Similarly, if a user is working on a deal linked to a specific Jobcode, that Jobcode will be of far greater relevance during the deal than six months after the deal has finished. The chart icon in Figure 6 and those in Figure 7 can generate a chart showing the evolution of importance over time – for a definable period set by the user. This popup is presented in Figure 8.

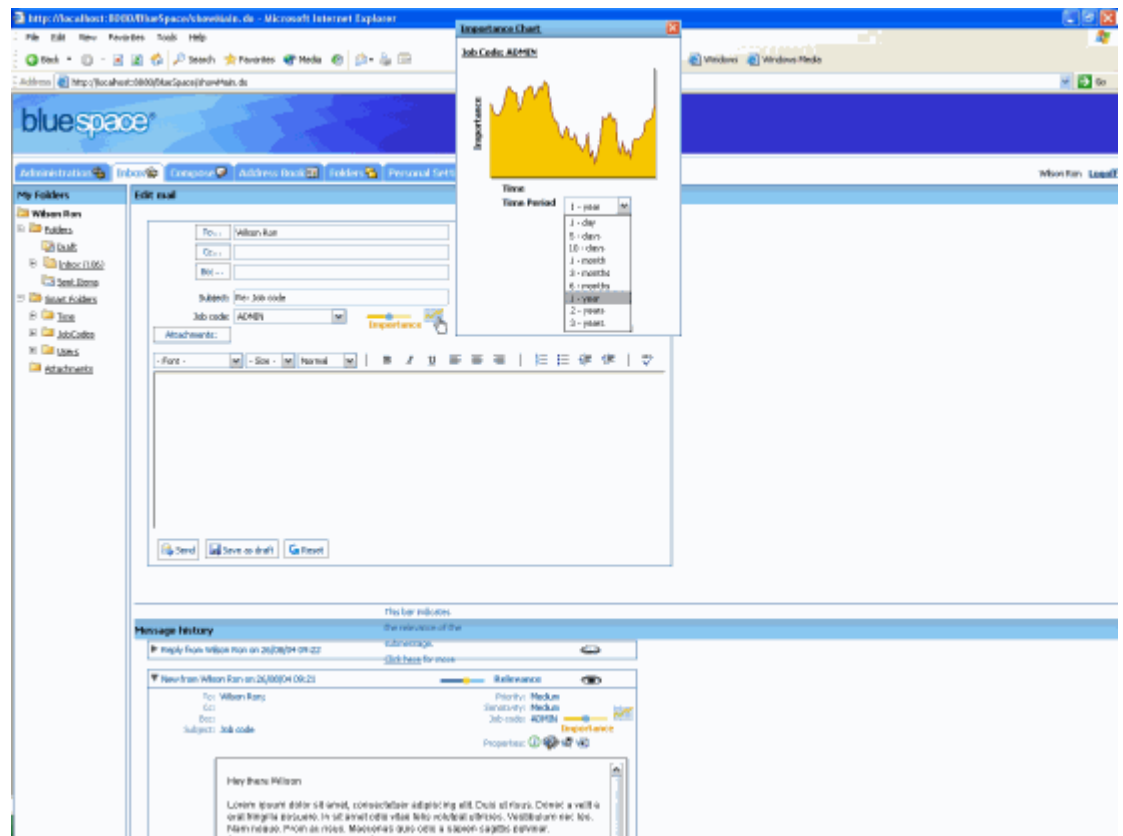


Figure 8 – Importance evolution

The evolution of importance over time is a critical concept in searching a set of TransMails, as discussed in detail in Section 7.3.

7.2.2 Folder view

The folder view is the way BlueSpace presents a TransMail folder listing to the user, showing select details of, say, 20 different TransMails at the same time. This is illustrated in Figure 9.

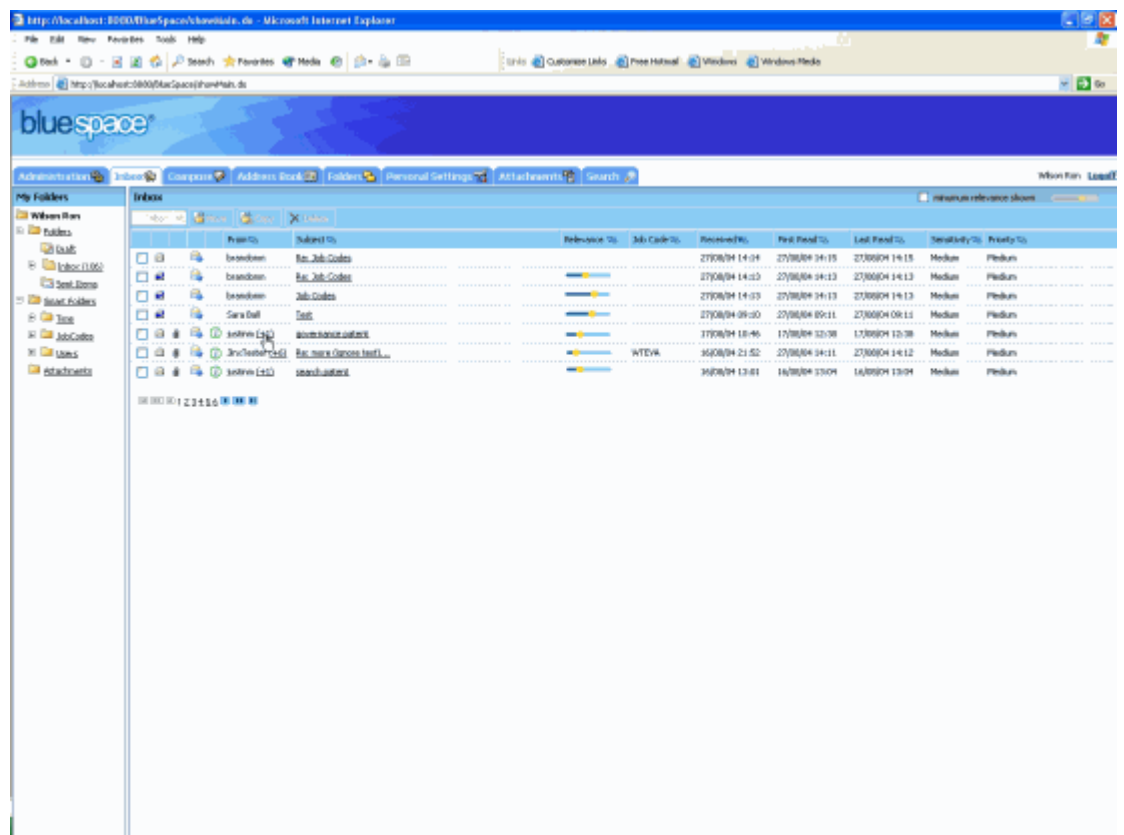


Figure 9 – BlueSpace TransMail folder view

The table in Figure 9 includes a column called relevance, which shows a relevance slider for every TransMail in the folder listing. As in the mail view in Section 7.2.1, the user can move the relevance slider to provide feedback to the application. In addition, the user can click the details icon which presents the same new box as that shown in Figure 7, indicating users' actions and importances.

7.2.3 Address book

Figure 10 shows the BlueSpace address book. Each user (as well as each group and domain) are shown, and can be added to a user's favourites (those

they communicate with regularly).

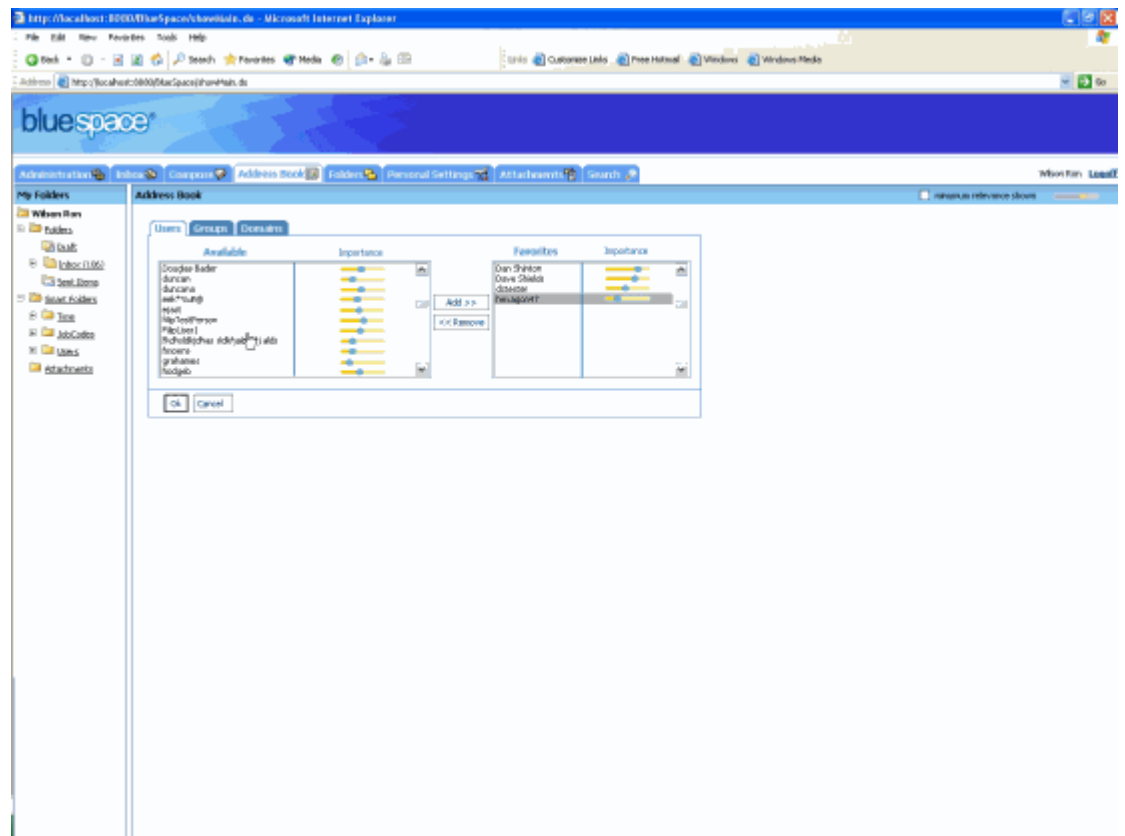


Figure 10 – BlueSpace address book

Figure 10 also shows an importance slider for each user, group or domain, and allows the user to directly interact with the importance values given to their contacts. Again, the user can either lock the value or provide it as a new start value from which the system equilibrates.

7.2.4 Preferences

Using preferences the user can affect the different inputs into the relevance algorithm. This allows the user to tune the relevance algorithm to provide the content they are interested in. They can affect the relative strength of the different actions by users, as well as affecting the relative strength of the different entities with importance (Jobcodes, users etc). This is depicted in Figure 11.

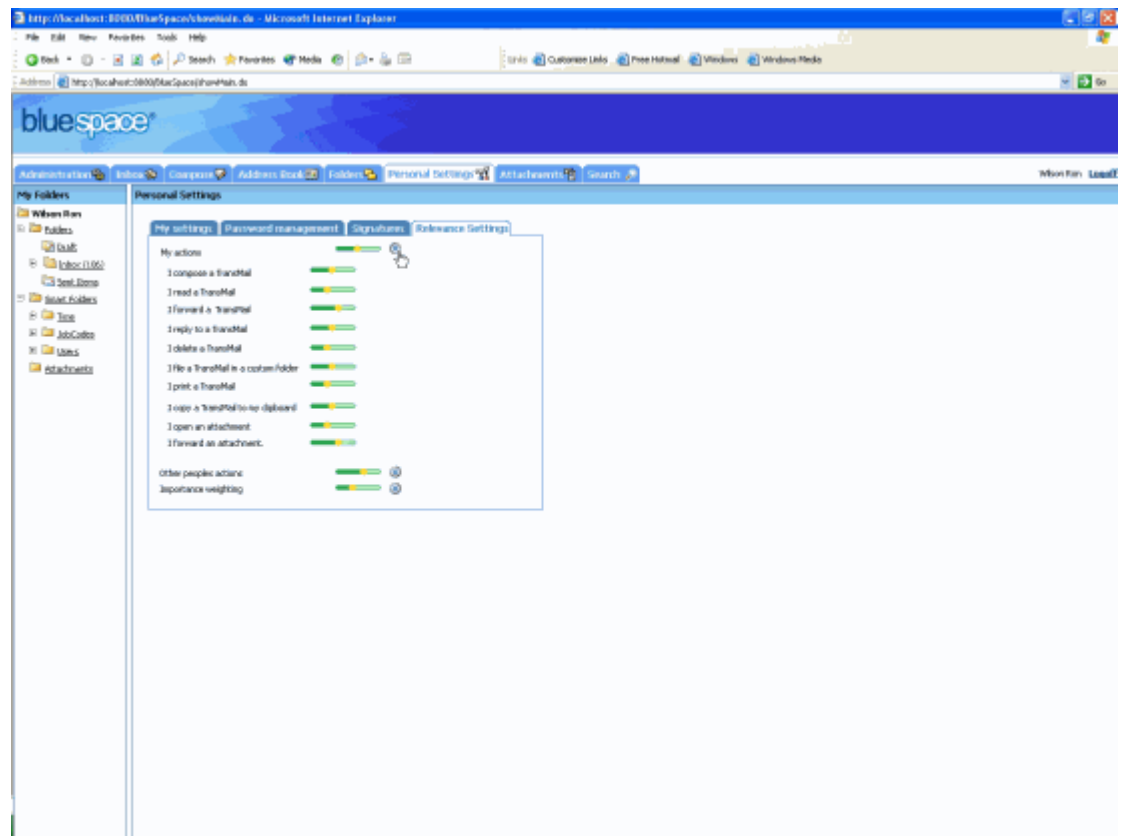


Figure 11 – Tuning the algorithm

In addition to varying the relative strengths of various inputs into the relevance algorithm, the user can also alter the importance of the different entities in the system, using an interface similar to that depicted in Figure 10.

Another option for searching is to display the simple search capability, as discussed in Section 1.1, on every page of the application, allowing the user a familiar search function from wherever they are in the application.

7.3 BlueSpace TransMail Search

BlueSpace TransMail search has been discussed extensively during this white paper, and it should be of little surprise that the search interface has far more flexibility and intelligence than any other mail search system. The search application can be divided into two areas – the search form and the search results.

7.3.1 Search form

The simple search form interface is presented in Figure 12.

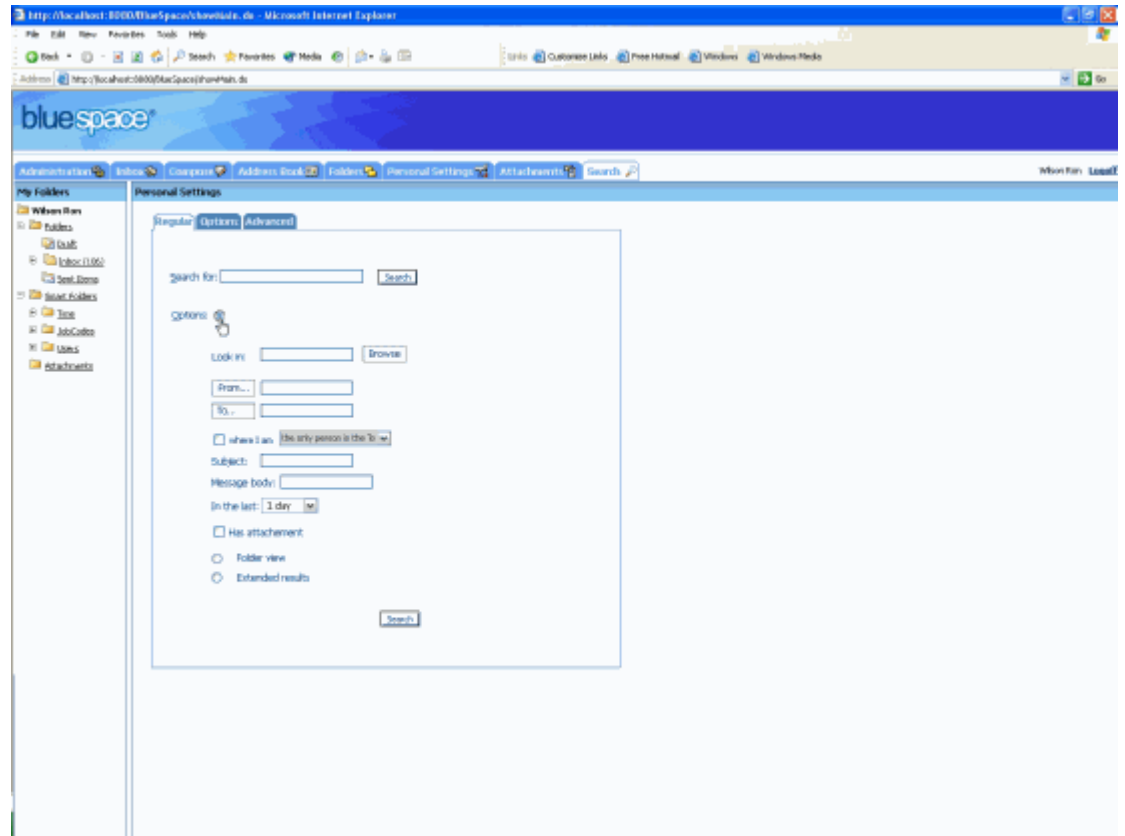


Figure 12 – Simple search interface

The tab shown is the simple search option, in which a user just enters any keywords they are interested in. The search algorithm then searches for this text in all possible criteria – users, groups, Jobcodes, message subjects, message bodies etc. This option provides an incredibly simple search algorithm for an average user.

If the user clicks the 'options' roll down in the search form interface shown in Figure 12, the application then provides some additional search options, as shown in Figure 13. Note that the application 'remembers' when the user clicks the options roll down, and serves the search interface to them with the options already showing the next time the user goes to the search page.

The options roll down allows the user to specify key words to search for,

Figure 13 – Normal search interface

Can allow a search for “mail I wrote that was forwarded to John”. Or “mail I wrote that became part of a tree linked to Jobcode X”.

Allow advanced tab for user to change the relevance sliders.

Need to allow the user direct interaction with relevance when searching, so allow them to use normal keywords etc, but also change the relevance sliders with respect to different parameters and update the search results.

Allow user to rate senders with a relevance score.

Relevance slider on your inbox – start with 20% and then increase it to try and narrow down you’re starting position when coming back from holiday.

Button for “why was this relevant?” to understand why the system picked it.

Need a button for “content like this” as per Google.

“What am I trying to achieve?” when searching, might change the results a bit. So “Find someone’s name” or “Find historical document” or “Find information about a project”. Have different inter-result correlative profiles.

Concept of being able to show results over time – so for example, plot a graph showing the frequency of results related to Jobcode X, will indicate activity on it. Also, if you remember that the comment you are looking for was made in the intensive period, then you could drag vertical lines in to the data graph saying “search between these dates”.

Entity diagrams – visualise the correlative strength between different entities (eg users/Jobcodes/groups) and even how these evolve over the lifecycle of a project. Management consulting inputs, but also useful for the end user.

This is also related to search. Say I knew that I wanted something to do with the Capita project but couldn’t remember the name of the project manager. Could use entity diagram to ‘see’ the project Capita and then see that Jason was very correlated to it somehow using relevance. And then click Jason to see the other projects that he was involved in.

Could allow results by different relevance weightings – so tabs in the search page for Jobcode/Date/User/Context etc. This affects the search results.

Should be able to search within search results, and should have a back button to get back to the previous search results set.

8 Algorithms

8.1 Brain Dump

We need to compare to Google's page rank technology, and there's a great study of it here: <http://www.webworkshop.net/pagerank.html>

Other links:

<http://ausweb.scu.edu.au/aw2k/papers/lowe/paper.html>

Need an expert in this area.

Calculate relevance for content as it comes in, then change it each time a user does an action, then weight it for people relevant to me, then recalculate importance scores for people on say nightly basis.

May be that due to outliers and random behaviour, if one person leaves a mail open for 30 minutes it's less than 10 times relevant than if ten people do. Need a way to deal with outliers – but not likely to get more than one.

If you forward to your boss and that's more important than one you sent to the office secretary, so you need to run the algorithm iteratively. This is because if the boss is of higher importance the secretary, then forwarding a message to the boss from Bob increases Bob's importance more than if you forwarded it to the secretary.

Importance of content can be done in realtime, importance of people can be reassessed each night.

Taking lag time – so longer you wait before reading, the less important.

Most others deleting mail – might want to weight the users again.

Different people can have different relevance ratings for different respective users – my boss matters more to me than they do to someone else in a different department. So need an n by n matrix.

Each user can affect the relevance scale for any other user as part of a learning cycle / feedback loop.

Big difference compared to normal search engines – each user gets different results.

Relevance scores need to drop off over time – so data that's two years old should have less impact than stuff that's 2 months old.

Relevance actually varies over time – during a particular date period a user may be much more relevant to me due to the project I was working on with them. This clearly reflects search, as relevance could be based on today's relevance values or the relevance value relative to the date the content was written.

Searches should go back into feedback loop – potentially has greater impact on future searches than on messages I'm interested in now, as relevance criteria changed over time. What you're interested in when searching may be very different to what you're interested in when reading – Dilbert cartoons may be the highlight of your day, but you never search back to find particular ones.

Multi-dimensional – if I email Andrew cc Yul and he always keeps it, then email Andrew cc Dan and he always deletes it, then different relevance.

Need to think about common themes in searches – how do you delve into this. Have all the data in a structured way, but how do you recognise themes and related themes through quasi-structured mail content.

Note that in the TransMail view, the TransMail relevance should be some algorithm summing the relevance of the submessages (both current and history) though the relevance components should drop off over time. So the further back in the history you go, the less effect that submessage has on the TransMail's overall relevance.

8.2 Initial Maths

NOTE THAT THIS INITIAL MATHS HAS BEEN DRAFTED WITH A MATHEMATICIAN, BUT IS IN NEED OF SIGNIFICANT CHANGES. IT IS INCLUDED HERE FOR COMPLETION IN THIS CURRENT DRAFT.

8.2.1 Introduction

There are three lists of attributes which could feasibly influence the Transmail's relevance to a receiver user. The three lists are groupings according to mail actions relevance, collaborative relevance and defined importance factors, which I shall label A , B and C respectively. The most obvious starting point is to work with indicator functions, $J(\cdot)$, which are simply "on" or "off", depending on whether the attribute is favourable or not. This is fine for attributes such as:

"If a user receives a TransMail and replies to it, then the content of the original TransMail is likely to be of significant relevance",

but not necessarily for attributes such as:

"If a user has read a given TransMail many times, then it is likely to be of significant relevance to them.",

where a score function may be more appropriate, with a higher score dependent on the number of times opened or read. However this causes problems of relative weighting amongst attributes, and in any case we can create an indicator function by defining a threshold value of number of times opened or read. The third list is different as it deals with importances (see later), and thus must have a continuous scale of values. These importance functions are called $I(\cdot)$. A working formula for the relevance, $R(T_M)$, of a given transmail, T_M is then:

$$R(T_M) = \frac{1}{3} \left(\frac{1}{W_A} \sum_{j=1}^{n_A} w_j^A J(A_j) + \frac{1}{W_B} \sum_{j=1}^{n_B} w_j^B J(B_j) + \frac{1}{W_C} \sum_{j=1}^{n_C} w_j^C I(C_j) \right)$$

where n_A , n_B and n_C are the number of attributes in each list, w^A , w^B and w^C are weightings associated with each attribute to allow for the possibility that some attributes are deemed of greater intrinsic relevance than others, and W^A , W^B and W^C are the sum of the weightings in each of the three lists. Of course if the weightings are all unity in a list then $W = n$ for that list. This formula assumes that the relative maximum weighting of each list is equal to 1, a sensible option; it may also be sensible to weight each list in the ratio $W_A : W_B : W_C$, in which case the formula would be:

$$R(T_M) = \frac{1}{W} \left(\sum_{j=1}^{n_A} w_j^A J(A_j) + \sum_{j=1}^{n_B} w_j^B J(B_j) + \sum_{j=1}^{n_C} w_j^C I(C_j) \right)$$

where:

$$W = \sum_{j=1}^{n_A} w_j^A + \sum_{j=1}^{n_B} w_j^B + \sum_{j=1}^{n_C} w_j^C ,$$

the sum of the weight within and across the lists.

8.2.2 Importance

The third list is special in that it introduces the interactive aspect of relevance between users through entity importance, whereas the first two are purely independent in their individual factor contributions to relevance. This is because the intrinsic nature of these factors for mail actions by the user or by other users other than the user in question will affect the relevance of the Transmail, but the relevance of the Transmail will not feedback into the quantification of these factors. They either happen or they don't – the fact that the relevance may dictate whether these actions are engaged is irrelevant here, as we are merely revising relevance in the light of **past** actions. However, the importance factors are different. Here the values of these factors are affected by the email traffic that has occurred. For example the importance of a user is affected by the receipt of Transmails from other users – his/her importance perturbs or is shared by the importance of the other users. This interactive process must be taken into account before the importance can be fed into the equations (1) or (2) in a search. Similar arguments will apply for other importance factors, so we will assume that we can focus on one such importance factor, say the user importance factor.

To estimate the importance of a user the following model can be considered.

1. There is, for user i , an intrinsic immutable importance, I_i^0 , which is set globally and presumably takes into account levels of responsibility within the organisation. This intrinsic value should ensure that the mechanism doesn't result in total uniformity of user importance.
2. If a Transmail is sent from user i to user j , the importance of user j is amended slightly such that if user i is of higher importance than j then the importance of j is raised; however if the importance of i is lower than j then the importance of j is lowered. There is in essence a sharing of importance from the sender to the receiver. The importance of the sender is unaltered by the Transmail, but of course his/her importance **relative** to the receiver has been altered, such that the difference between them has been reduced.
3. The "sharing parameter", α , dictates the degree of this sharing, with $0 < \alpha < 1$. In practice α will probably high, say 0.9 or higher. There is also a parameter β to allow for the restraining to the preset level, I_j^0 .

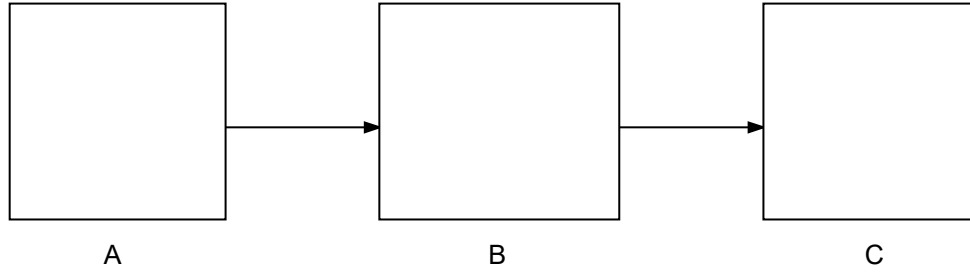
With this, an equation to update user j 's importance at the k th update from Transmail's received since the last update, $k-1$ is:

$$I_j^{(k)} = \beta I_j^0 + (1 - \alpha - \beta) I_j^{k-1} + \frac{\alpha}{m} \sum_{i=1}^m I_i^{k-1} / C_i$$

where C_i is the number of recipients of the i th email. Note in the summation the same user may appear more than once. Note unless we divide the third term in the formula above by m , we will always increase the importance of a user by receipt of Transmails. This should not, of course, be the case.

Furthermore if no emails are received the importance is slightly reduced; this is feasible as it indicates a nonparticipatory aspect of the user.

8.2.3 Example



User A has sent a Transmail to B who has sent a Transmail to C. Here:

$$I_A^k = \beta I_A^0 + (1 - \alpha - \beta) I_A^{k-1}$$

$$I_B^k = \beta I_B^0 + (1 - \alpha - \beta) I_B^{k-1} + \alpha I_A^{k-1}$$

$$I_C^k = \beta I_C^0 + (1 - \alpha - \beta) I_C^{k-1} + \alpha I_B^{k-1}$$

This system of equations will be solved iteratively in what is essentially a Gauss-Seidel algorithm procedure until convergence to a sensible precision. This procedure will then be repeated every regular time period, say daily. In this example, the more important A is, the more important B becomes, which then increases C's importance more than if B had not received the email from A.

The importance model above allows for relative importance for B with respect to A, as the adjustment resulting from a Transmail between A and B depends on the relative values for I_A and I_B .

8.2.4 Time-dependent relevance

Returning to the main framework equation for relevance (for example equation (1)), we would like a received Transmail's relevance to generally a time-dependency. We should therefore premultiply the formula in equation (1) by a sensible function of time, $G(t)$:

$$R(T_M) = \frac{G(t)}{3} \left(\frac{1}{W_A} \sum_{j=1}^{n_A} w_j^A I(A_j) + \frac{1}{W_B} \sum_{j=1}^{n_B} w_j^B I(B_j) + \frac{1}{W_C} \sum_{j=1}^{n_C} w_j^C I(C_j) \right)$$

For example $G(t)$ might reflect relevance of a project up to a deadline time, t_0 , beyond which the relevance begins to decay:

$$G(t) = \begin{cases} 1 & t \leq t_0 \\ e^{-\lambda(t-t_0)} & t > t_0 \end{cases}$$

where λ is a decay constant (high for rapid decay, low for slow decay).

8.1.5 Notes

1. These formulae are additive in their nature between the three lists, but for importance is interactive within. A further degree of sophistication might be to include degrees of effect, say on user action which is dependent on the importance of the user or other entities. These are then interactions between the actions and the importances.
2. There is no effect on the sender's importance of the Transmail in this model. We could model alternatively with the adjustment to importance being made to the sender rather than the receiver, or both.
3. The model for importance is reminiscent of that used for the PageRank system by Google, but is not exactly the same. The same iterative procedure, which will employ the basic Gauss-Seidel algorithm is employed however.

9 Conclusion

We are brilliant!!!

-
- ¹ [Talking Headers](#), Katie Hafner and Matthew Lyon, The Washington Post Magazine, 4th August 1996.
- ² [Simple Mail Transfer Protocol](#), Johnathan Postel, RFC 821, August 1982
- ³ [Legal Lottery](#), Melanie Liew, Computerworld, 28th April 2004
- ⁴ [Talking Headers](#), Katie Hafner and Matthew Lyon, The Washington Post Magazine, 4th August 1996.
- ⁵ [Simple Mail Transfer Protocol](#), Johnathan Postel, RFC 821, August 1982
- ⁶ [Is banning e-mail a viable way for IT to avoid being crippled by storage costs?](#), Julia Vowler, ComputerWeekly, 16th October 2003
- ⁷ [Email management too important for CIOs](#), Rachel Fielding, Network News, 11th February 2003
- ⁸ [Legal Lottery](#), Melanie Liew, Computerworld, 28th April 2004
- ⁹ *Relational Messaging White Paper*, BlueSpace Group Ltd, 2nd July 2002
- ¹⁰ *BlueSpace Discovery Service*, BlueSpace Group Ltd, 3rd December 2003
- ¹¹ *BlueSpace Email Structuring*, BlueSpace Group Ltd, August 2004
- ¹² *BlueSpace Integrated Governance Management*, BlueSpace Group Ltd, 15th May 2004
- ¹³ *BlueSpace Integrated Governance Management*, BlueSpace Group Ltd, 15th May 2004
- ¹⁴ *BlueSpace Context*, BlueSpace Group Ltd, August 2004
- ¹⁵ *BlueSpace Email Structuring*, BlueSpace Group Ltd, August 2004
- ¹⁶ *BlueSpace Integrated Governance Management*, BlueSpace Group Ltd, 15th May 2004

EXHIBIT 2004

metadata slip sheet

BegDoc	REV0000100
BegAttach	REV0000100
File Name	BlueSpace Mail Search.doc
LastCreateOn	8/4/2004 1:48 PM
LastModifiedDate	9/2/2004 1:23 PM
Custodian	Marston, Justin
Author	.
SortDate	9/2/2004 1:23 PM