Análisis de listas de correo en el software libre: un caso de estudio

Israel Herraiz, Juan José Amor, Álvaro Navarro

Resumen— **Prácticamente todas las comunidades de software libre (proyectos, asociaciones de usuarios, etc) tienen una o más listas de correo electrónico. La organización, el intercambio de información y las decisiones relativas a la comunidad se toman en gran medida en estas listas de correo-e, de manera que todos los miembros puedan intervenir o al menos permanezcan informados de lo que acontece. No es difícil, por tanto, inferir que el nivel de actividad y la participación en las listas de correo de una comunidad son un indicador del nivel de actividad de ésta. Partiendo de esta hipótesis, hemos creado una herramienta llamada MLStats que nos permite obtener datos de participación de listas de correo. Hemos utilizado la lista de socios de la asociación Hispalinux, una de las agrupaciones de GNU/Linux más grandes a nivel mundial, como caso de estudio de nuestra herramienta. Los resultados obtenidos muestran que en el último año la participación y la actividad en esta lista de socios está disminuyendo.**

I. Introducción

Las comunidades de software libre (proyectos, grupos de usuarios) se encuentran dispersas geográfica y administrativamente [1]. El surgimiento de Internet ha posibilitado el nacimiento de comunidades dispersas geográficamente, en las que a menudo sus miembros no se conocen en persona, a pesar de haber estado años interactuando a través de Internet. El fenómeno del software libre es, junto a otras comunidades, pionero en este sentido [2]. De hecho, comunidades científicas están estudiando desde hace poco los fenómenos derivados de la mencionada dispersión [3]. Uno de los medios, aunque no el único, que hace posible una ágil comunicación entre miembros de equipos de trabajos dispersos son las listas de correo electrónico.

El funcionamiento de las listas de correo se basa en el intercambio de correos electrónicos entre los miembros subscritos a dicha lista. Una persona puede iniciar una conversación enviando un mensaje a la lista siendo éste recibido por todos los miembros de la lista en el mismo instante de tiempo. Posteriormente, cada usuario comprobará su correo electrónico y enviará una respuesta si lo cree oportuno.

HispaLinux es una de las asociaciones de GNU/Linux más grande a nivel mundial (si no la más grande) que ha experimentado un rápido crecimiento en los últimos años, llegando a contar con más de 7000 socios, con una alta dispersión geográfica entre sus miembros activos; y coordina muchas de sus actividades a través de sus listas de correo, especialmente la lista de socios, en la que cualquier miembro de la asociación puede participar libremente. Gracias al análisis que proponemos, podremos comprobar si la actividad en la lista de socios ha sido acorde con el crecimiento experimentado o si, por contra, no ha sido así. Por supuesto, la serie de factores que afectan a la participación puede ser múltiple y no sólo dependen del número de socios, incluyendo aspectos organizativos, políticos, etc. Por otro lado, las características de la asociación, en número y dispersión de sus miembros, hacen de la lista de socios un interesante caso de estudio del análisis de listas de correo-e.

II. La asociación Hispalinux

Hispalinux es la asociación española de usuarios de software libre (en especial GNU/Linux), que cuenta en la actualidad con más de 7000 socios. Cualquier persona que cumpla los requisitos legales del asociacionismo en España puede hacerse socio de Hispalinux, mediante un formulario que se encuentra disponible en su página web [4].

La asociación fue fundada en junio de 1997, aunque su fundación se gestó durante los dos años anteriores. Los fines de Hispalinux han ido siempre ligados a la divulgación del software libre en España y el mundo hispanohablante, y fueron numerosas las personas que colaboraron con entusiasmo en sus primeros pasos.

Algunos de los miembros fundadores de Hispalinux estaban ligados a otros proyectos exitosos de los usuarios hispanohablantes, específicamente el proyecto Linux en Castellano, LuCAS [5]. La aparición de la asociación sirvió para que estos proyectos pasaran a disfrutar de los recursos informáticos que la asociación proveería.

Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos

{herraiz, jjamor, anavarro}@gsyc.escet.urjc.es

Paralelamente a este apoyo a otros proyectos de la comunidad, la asociación ha emprendido numerosas actividades divulgativas, destacando las relacionadas con el impulso del software libre en la administración pública española [6].

En la actualidad, Hispalinux tiene una apariencia más *profesional* o al menos más alejada del aspecto lúdico que tenía hace unos años. Hay diversos grupos de trabajo especializados en diversas áreas. Por ejemplo, hay voluntarios dedicados al gabinete de prensa, otros dedicados a la administración de sus servidores y otro específico para el portal `software-libre.org` [7], donde se alojan los proyectos de software libre de los socios, y de los miembros de las empresas y organizaciones que tienen un acuerdo con Hispalinux. Una característica básica es que los grupos de trabajo suelen estar coordinados a través de una lista de correo, por lo que la actividad en éstas pueden ser buenos indicadores de la vitalidad de cada grupo, o de la asociación en su conjunto.

Otra de las actividades principales de la asociación, que viene celebrando desde sus inicios, es su congreso anual. Iniciado en 1998, la tercera edición de 2000 reunió a más de 1000 personas y contó con la participación de muy destacados ponentes. En 2004, sin embargo, no se celebró, por primera vez tras seis años de éxito.

A finales de 2003 y principios de 2004 se produjeron algunos hechos relevantes de la asociación: en octubre de 2003 dimitía de su cargo el entonces secretario de la asociación, coautor de este artículo; y en julio del año siguiente se celebraron elecciones a Junta Directiva tras la dimisión en bloque de ésta. La Junta Directiva fue reelegida (al menos dos personas repiten en los cargos de presidente y vicepresidente), y continúa al frente de la asociación.

III. La herramienta MLStats

La primera versión de MLStats se desarrolló como parte de un proyecto Fin de Carrera dentro del Grupo de Sistemas y Comunicaciones de la Universidad Rey Juan Carlos, y fue escrita por Javier Crespo [8][1].

La versión empleada para este artículo ha sido reimplementada desde cero en lenguaje Python[2], usando el paradigma de programación orientada a objetos. La figura 1 muestra la arquitectura de la herramienta. Consta de seis clases: una clase principal que representa a la aplicación, un *parser* de HTML, un generador de informes (con el que se crearon las gráficas usadas en este artículo), un *parser* de correo electrónico (usando el *estándar* RFC822[9]), y una clase para representar base de datos genéricas (de la que se heredó otra clase con los detalles de la base de datos referentes a la aplicación).

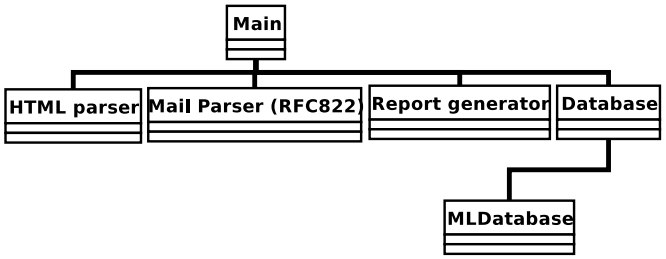

Fig. 1. Arquitectura de MLStats.

El procedimiento empleado por la herramienta es el siguiente:

- Descarga la página web donde se muestra el resumen del archivo de una lista de correo.
- Realiza un análisis del código HTML de esa página en busca de enlaces de un determinado tipo (por ejemplo, `txt.gz`).
- Descarga todos los enlaces encontrados.
- Desempaqueta si es necesario cada fichero.
- Analiza cada fichero, según el estándar RFC822, e introduce cada nuevo mensaje analizado en la base de datos.
- A partir de la base de datos, calcula algunas estadísticas de la lista de correo, y genera algunas gráficas.

En el caso de este artículo, se obtuvo un fichero *mbox* con el archivo completo de la lista. El fichero fue proporcionado tras una petición en la propia lista de socios de Hispalinux. Por tanto, no se realizaron los dos primeros pasos mostrados en la enumeración anterior.

IV. Resultados

El archivo completo de la lista ocupaba 65 megabytes. Contenía 12.820 mensajes, de los cuales uno no había sido formado conforme al *estándar* RFC822, por lo que fue descartado. El primer mensaje se envió el 6 de noviembre de 2000 mientras que el último mensaje analizado data del 27 de septiembre de 2005[3]. El mes con más mensajes fue septiembre de 2003 (que coincide con la celebración del congreso anual). El mes con menos actividad fue agosto de 2005.

El primer análisis que hemos realizado para comprobar la evolución del nivel de actividad de la lista de

[1]La última versión, ligeramente modificada y corregida, está disponible bajo licencia GPL en http://libresoft.urjc.es/index.php?menu=Tools&Tools=MLStats

[2]http://www.python.org

[3]Nótese que no se dispone de los mensajes de toda la historia de la asociación, debido a que con anterioridad no se archivaban los mensajes de la lista

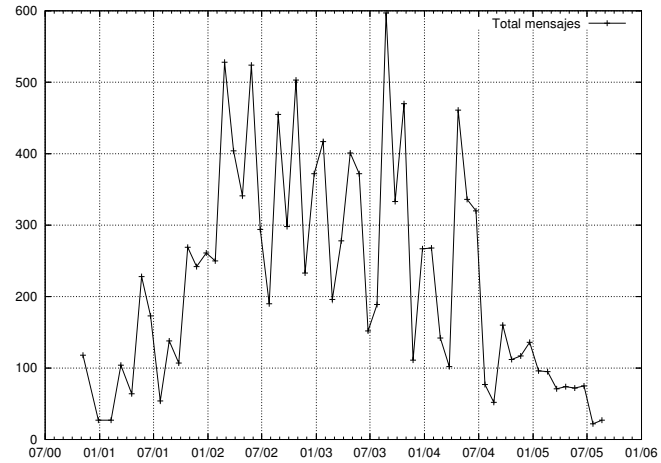

Fig. 2. Número de mensajes enviados a la lista cada mes.

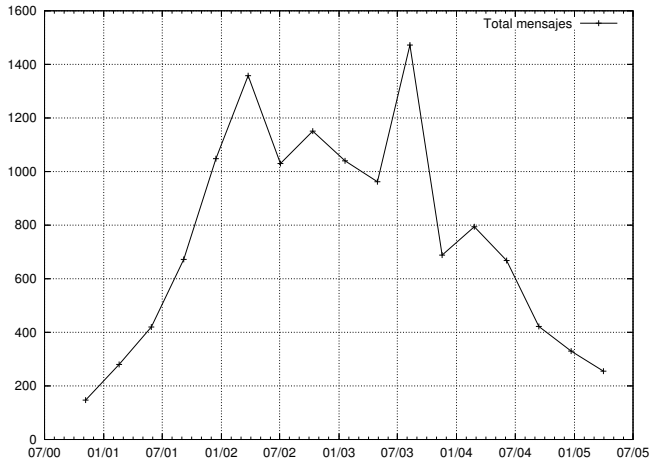

Fig. 3. Número de mensajes enviados a la lista cada 100 días.

correo ha sido calcular el número de mensajes enviado cada mes a la lista. Los resultados se muestran en la figura 2, en cuyo eje vertical se representa el número de mensajes y en el horizontal la fecha (en formato mes/año).

En la figura 2 se observan varios fenómenos. En primer lugar, parece claro que el apogeo de actividad en la lista ya ha pasado, y que se produjo en los años 2002 y 2003. Además, se observa también la estacionalidad en la actividad en la lista; esto es, en los períodos vacacionales (alrededor de diciembre y agosto de cada año), la actividad es mucho menor que en el resto de meses para cada año. Pero el hecho más importante es que desde finales de 2004 la actividad ha descendido drásticamente, y en estos momentos está a niveles similares a los del comienzo de la lista, y con tendencia a la baja. Esto parece evidenciar que la participación en la lista ha disminuido.

Algunos de los picos de la figura 2 (octubre a noviembre de 2003) parecen coincidir con los momentos en los que se produjo la dimisión del secretario. Tras este momento, la actividad disminuyó bruscamente, y se produjo la dimisión en bloque de la Junta Directiva en marzo de 2004. Parece que tras esta dimisión en bloque, la actividad se recuperó durante algunos meses (en abril de 2004 se fundó `software-libre.org` [7], en mayo comenzaba la campaña electoral, en julio se celebraron las elecciones), aunque finalmente durante este último año ha sido mucho menor que en años anteriores.

Para filtrar los picos producidos por la estacionalidad, podemos tomar períodos de tiempo mayores entre cada dos puntos. Por ejemplo, con períodos de 100 días, la figura 2 se transforma en la figura 3. Las conclusiones son las mismas que tomando períodos de tiempo más cortos.

Para comprobar cómo ha evolucionado el nivel de participación, hemos calculado el número de personas que han participado en la lista cada mes. Para medir el número de personas hemos tomado dos métricas diferentes: número de nombres completos diferentes, y número de direcciones de correo electrónico diferentes. En total, durante toda la historia de la lista hemos contabilizado 1197 nombres diferentes y 1134 direcciones de correo electrónico diferentes. Si tenemos en cuenta el campo `From` completo (esto es, nombre y dirección de correo electrónico), encontramos 1646 participantes diferentes en la lista. Como una persona puede escribir su nombre de diferentes modos (por ejemplo, abreviando el segundo nombre), y puede emplear diferentes direcciones de correo electrónico, estos datos suponen una cota superior al nivel real de participación. Además, es evidente que combinando nombre y dirección de correo electrónico, la cota obtenida es más burda; consideremos el caso por ejemplo de una persona que escribe siempre con el mismo nombre, pero con diferentes direcciones de correo electrónico.

El mes con más participantes fue noviembre de 2002, y el mes con menos participantes fue agosto de 2005 (independientemente del criterio usado para discriminar entre dos participantes diferentes).

En cuanto a la evolución de la participación, en la figura 4 se muestra el número de personas que han escrito al menos un mensaje durante cada mes, usando el nombre como criterio de discriminación entre personas (la gráfica diferenciando por correo electrónico es muy similar).

Los mayores niveles de participación coinciden con los momentos de mayor actividad en la lista. El fenómeno de estacionalidad es menos acusado en el nivel de

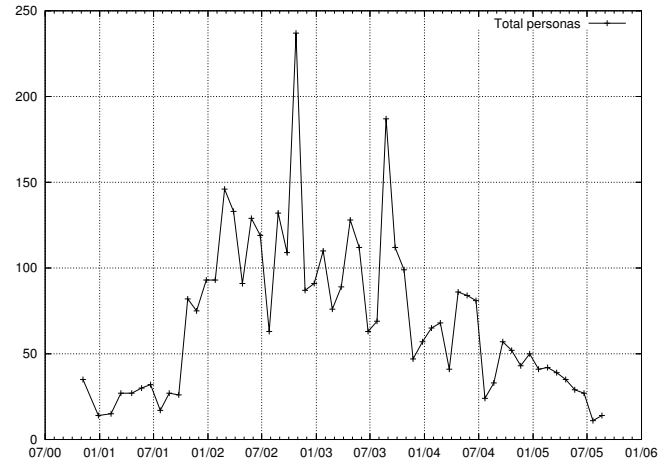

Fig. 4. Número de personas que participan en la lista cada mes (diferenciando por nombre).

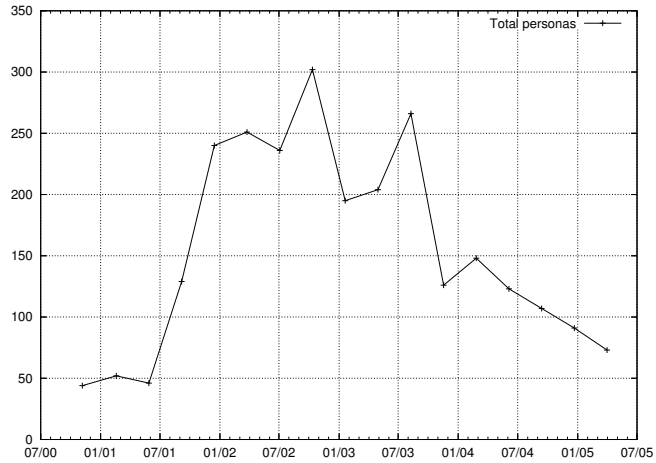

Fig. 5. Número de personas que participan en la lista cada 100 días (diferenciando por nombre).

participación, lo que indica que en los períodos vacacionales disminuye la cantidad de envíos que realiza cada persona en mayor medida que la cantidad de personas que participan en la lista. Es decir, existe cierta fidelidad a la lista en los períodos de vacaciones, pero con un menor número de envíos que en períodos no festivos. La observación más interesante es la disminución progresiva de la participación; tras un período de crecimiento inicial, el número de personas que participa en la lista ha disminuido gradualmente, hasta alcanzar niveles similares al comienzo de la lista.

Si nos fijamos en el período de octubre a noviembre de 2004 , parece que el comportamiento en la participación mimetiza a la actividad de la lista. Esto es, tras la dimisión del secretario (octubre de 2003) se incrementó algo la participación, para caer bruscamente poco después. A esto siguió la dimisión en bloque de la Junta Directiva (marzo de 2004), que de nuevo tuvo como efecto un ligero incremento en la participación. Sin embargo, estos niveles no han logrado mantenerse, y en el último año la participación ha disminuido.

De nuevo, podemos filtrar los efectos de la estacionalidad tomando períodos de tiempo mayores. En la figura 5 se muestra la evolución del número de personas (diferenciando por nombre) que escriben a la lista en períodos de 100 días. De nuevo, la gráfica diferenciando por dirección de correo electrónico es muy similar y no se ha incluido por razones de espacio. Independientemente del período de tiempo elegido, la disminución de la participación en la lista es evidente.

Otro síntoma de la salud de la actividad de la asociación es la concentración de mensajes por persona. Lo ideal es evitar la concentración de la participación, para que no aparezcan individuos cuya intervención sea fundamental para mantener el nivel de la actividad de la asociación. Para estudiar este aspecto hemos realizado diferentes análisis. Por un lado mostramos la distribución de la participación en la lista (de nuevo tomando nombres y direcciones de correo electrónico como criterios de diferenciación entre personas), y hemos calculado el coeficiente de regresión lineal entre el número de mensajes enviados a la lista cada mes y el número de personas (diferenciando por nombre y por dirección de correo electrónico) que envían mensajes a la lista cada mes.

La correlación lineal entre el número mensual de mensajes y de personas con diferente nombre que al menos escriben un mensaje arroja un coeficiente de $r = 0{,}9702$. En el caso de personas con diferentes direcciones de correo electrónico, el coeficiente es de $r = 0{,}9666$. Esto parece indicar que la participación está distribuida entre todos los participantes. Comprobemos la distribución de la participación; en la figura 6 mostramos la distribución de los mensajes debidos a personas con diferente nombre, donde en el eje vertical se muestra el porcentaje de mensajes acumulados respecto al total, y en el horizontal el número de personas[4]. Como podemos ver, 24 personas diferentes son las responsables del 50 % de los envíos. En la figura 7 se muestra la misma distribución, pero esta vez tomando como criterio diferenciador la dirección de correo electrónico. En este caso, 32 personas son las responsables del 50 % de los envíos.

Por tanto, podemos concluir que la relación entre la actividad y la participación en la lista es saludable, en el

[4]Se ordenaron las contribuciones de cada persona, y se sumaron en orden descendente de participación. Es decir, primero se muestra la contribución de la persona que envió más mensajes, después las contribuciones de las dos personas que enviaron más mensajes, y así sucesivamente.

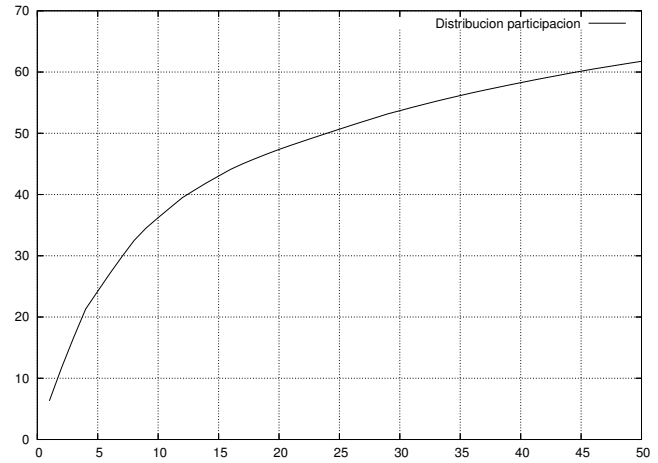

Fig. 6. Distribución agregada de los mensajes enviados por cada participante (discriminando por nombre).

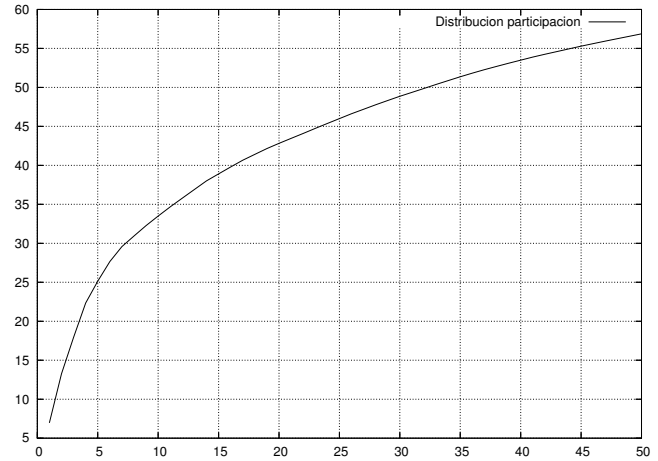

Fig. 7. Distribución agregada de los mensajes enviados por cada participante (discriminando por *e-mail*).

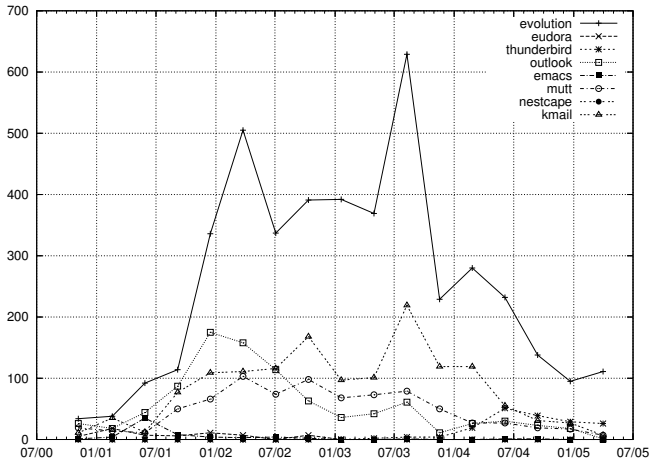

Fig. 8. Programas más usados (cada 100 días) en la lista.

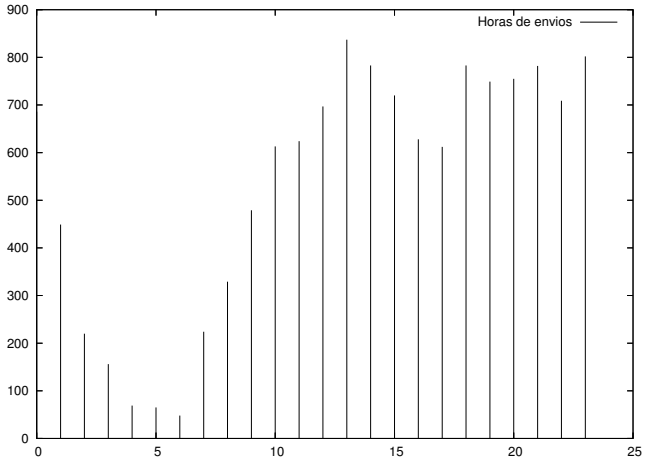

Fig. 9. Horas a las que se realizan los envíos

sentido de que no existen individuos fundamentales para mantener la actividad en la lista.

Otro dato que podemos obtener de la lista de correo es el programa cliente que se emplea para enviar el correo electrónico. En la figura 8 se muestra la evolución de una selección de los ocho programas más usados por los participantes en la lista de correo electrónico; en el eje vertical se muestra el número de mensajes, en el horizontal la fecha, y cada curva representa uno de los programas usados[5]; se ha contabilizado el número de mensajes enviados con cada programa cada 100 días. Como podemos ver, el programa más popular es Ximian Evolution, seguido de KMail. Destacamos también la popularidad de Microsoft Outlook así como Mutt, un clásico entre los entusiastas del software libre.

Por último, podemos analizar qué horas son las más frecuentes cuando se envían mensajes a la lista. Esta información se muestra en la figura 9, donde en el eje vertical se muestra el número de mensajes y en el horizontal la hora del día a la que se realiza (en formato 24 horas). Como podemos observar, los mensajes son más frecuentes alrededor de las 12 y las 20 horas, siendo mucho menos frecuentes durante la madrugada. Al ser Hispalinux una asociación de habla hispana, podría ser interesante comprobar cuántos hispanohablantes no españoles participan en las listas. La figura 9 parece indicar que la mayoría de los mensajes se producen en España, puesto que la hora mostrada es la local española. Es decir, los mensajes en horarios habituales de, por ejemplo, Latinoamérica se producirían en horas *no habituales* en España.

[5]Para contabilizar estos resultados comprobamos el contenido de los campos `user-agent` y `x-mailer` de la cabecera de los mensajes. Las curvas recogen todas las versiones de cada programa; por ejemplo, en el caso de Outlook se incluyen tanto Microsoft Outlook como Microsoft Outlook Express.

V. Conclusiones

Las listas de correo contienen información suficiente para estimar los niveles de actividad y participación de las comunidades de software libre. En el caso de este artículo, hemos estudiado la lista de socios de Hispalinux, el mayor grupo de usuarios de software libre de España. Esta asociación cuenta con socios en todas partes de España, por lo que una lista de correo es un medio excelente para mantener la comunicación entre un grupo de personas tan disperso.

Los datos muestran que la actividad está en niveles similares al comienzo de la lista de correo, y descendiendo. Así mismo, la participación está también disminuyendo, y está también en niveles similares a los del comienzo de la lista. Por el contrario, la actividad está bien distribuida entre los participantes, y no existen individuos cuya participación pueda considerarse fundamental para mantener el nivel de actividad.

La mayoría de los mensajes en la lista provienen de España, o al menos, se escriben en franjas horarias que podemos considerar horario habitual en España. Sería necesario intentar trazar el origen de los contribuyentes a la lista, teniendo en cuenta los dominios de sus cuentas de correo (obviamente, sólo en aquellos casos en los que no se empleen dominios internacionales como `.com`) para comprobar el origen geográfica de cada participante en la lista.

En cuanto a los factores que pueden haber provocado el descenso en la actividad, es complicado aventurar alguna hipótesis habiendo analizado sólo una lista. Parece que los momentos de mayor actividad coinciden con el congreso de 2003 y que los momentos de menos con el año en el que no se ha celebrado congreso; que los problemas internos en forma de dimisiones vinieron acompañados de una reactivación de la lista, pero no conocemos cómo estos problemas internos influyeron en los diferentes grupos de trabajo, ni si la celebración del congreso provoca un aumento de actividad o viceversa (y por tanto, si la ausencia de congreso se debe a la disminución de la actividad, o al contrario). Por tanto, sería necesario realizar un análisis global de todas las listas para intentar formular explicaciones más sólidas acerca del descenso de actividad en la lista de socios.

A pesar de la incompletitud de nuestro estudio, vamos a aventurar una de las posibles explicaciones del descenso de actividad. A nuestro juicio un factor importante ha sido la deslocalización y distribución de los grupos de usuarios. En los últimos años ha crecido el número de grupos de ámbito local o regional (por ejemplo, en el Libro Blanco del Software Libre en España [10] aparecen censados 162 grupos), que también organizan actividades, difunden el software libre, etc. Esto podría haber producido el desplazamiento de la actividad de los voluntarios a los grupos más cercanos geográficamente. Por tanto este descenso en el grupo de usuarios más significativo de España no tiene por qué significar una mala salud del software libre en España. Para arrojar más luz en este punto sería interesante comparar el tráfico y la participación en las listas de Hispalinux con el tráfico en las listas de diferentes grupos locales.

VI. Agradecimientos

Agradecemos la colaboración de la Asociación Hispalinux, que nos proporcionó el archivo de la lista de correo de socios.

Este trabajo ha sido financiado en parte por la Comisión Europea, bajo el contrato CALIBRE CA del programa IST, número 004337, por la Universidad Rey Juan Carlos bajo el proyecto PPR-2004-42 y bajo el proyecto CICyT número TIN2004-02796.

Israel Herraiz ha sido financiado por la Consejería de Educación de la Comunidad de Madrid y el Fondo Social Europeo (F.S.E.), bajo la beca 01/FPI/0582/2005.

Referencias

[1] Yuwan Ke; Kumiyo Nakajoki; Yasuhiro Yamamoto; Kouichi Kishida, "The co-evolution of systems and communitties in free and open source software development," in *Free/Open Source software development*, Stefan Koch, Ed. 2005, Idea Group Publishing.

[2] Daniel German, "The GNOME project: a case study of open source, global software development," *Journal of Software Process: Improvement and Practice*, vol. 8, no. 4, pp. 201–215, 2004.

[3] James D. Herbsleb, Audris Mockus, Thomas A. Finholt, and Rebecca E. Grinter, "An empirical study of global software development: distance and speed," in *ICSE '01: Proceedings of the 23rd International Conference on Software Engineering*, 2001, pp. 81–90.

[4] Asociación Hispalinux, "Página web," Oct. 1997, http://www.hispalinux.es.

[5] Asociación Hispalinux, "Proyecto LuCAS," Oct. 1997, http://lucas.hispalinux.es.

[6] Asociación Hispalinux, "Campaña software libre en la Administración," Oct. 1997, http://www.hispalinux.es/SLAdmon.

[7] Asociación Hispalinux, "Portal Software-Libre.org," 2003, http://software-libre.org.

[8] Javier Crespo Martín, "MailingListStat: Herramienta de análisis de archivos de listas de correo," Tech. Rep., Universidad Rey Juan Carlos, 2005, Proyecto Fin de Carrera (I.T.I de Sistemas).

[9] "RFC 822 - Standard for the format of ARPA Internet text messages," http://www.faqs.org/rfcs/rfc822.html.

[10] A.Abella; J. Sánchez; M.A.Segovia, "Libro blanco del software libre en España," 2005, http://www.libroblanco.com.