รายงานวิจัยฉบับสมบูรณ์

โครงการ

การพัฒนาการใช้โปรแกรม R ในการวิเคราะห์ข้อมูล

R Program Development for Research Utilization

โดย

หัชชา ศรีปลั่ง และคณะ

เดือน ปี ที่เสร็จโครงการตามสัญญา

ตุลาคม 2549

บทคัดย่อ

ปัจจุบันการวิเคราะห์ข้อมูลทางสถิติ ต้องอาศัยโปรแกรมสำเร็จรูปที่สามารถรองรับข้อมูลขนาดใหญ่และระเบียบวิธีทางสถิติที่ซับซ้อนได้ แต่ซอฟต์แวร์ที่มีประสิทธิภาพดังกล่าวส่วนใหญ่เป็นซอฟต์แวร์ทางการค้าที่มีราคาสูง แต่ในระหลังนี้ มีผู้พัฒนาโปรแกรม R ขึ้นมา โดยเป็นโปรแกรมที่มีสมรรถนะสูง เป็นที่นิยมในหมู่นักวิชาการและนักสถิติระดับโลก และคุณสมบัติที่โดดเด่นของมันคือเป็นซอฟต์แวร์ที่มีลิขสิทธิ์แบบ GPL คือใช้งาน แจกจ่าย และปรับปรุงพัฒนาได้ฟรี โดยไม่ต้องเสียค่าใช้จ่าย แต่ซอฟต์แวร์นี้ยังเป็นที่รู้จักและมีผู้ใช้เป็นน้อยมากในประเทศไทย และยังมีข้อจำกัดที่สำคัญในเรื่องเอกสารประกอบการใช้งาน เนื่องจากตัวโปรแกรมเองเน้นการเป็นภาษาคอมพิวเตอร์ ทำให้ผู้ใช้ทั่วไปเรียนรู้ได้ยาก

วัตถุประสงค์

โครงการนี้จึงมีวัตถุประสงค์เพื่อพัฒนา เผยแพร่ และสนับสนุนให้มีการใช้โปรแกรม R ซึ่งเป็นซอฟต์แวร์ทางสถิติที่ไม่ต้องเสียค่าลิขสิทธิ์ ในการวิเคราะห์ผลงานวิจัยทั้งทางด้านวิทยาศาสตร์สุขภาพและระเบียบวิธีทางสถิติ

การดำเนินงาน

การดำเนินงานเพื่อให้บรรลุวัตถุประสงค์ดังข้างต้น จึงมีสามกลุ่มกิจกรรมคือ 1. การพัฒนาการใช้โปรแกรม R ให้ใช้ได้ง่ายขึ้น 2. พัฒนาส่วนติดต่อกับผู้ใช้เป็นภาษาไทย และ 3. การจัดกิจกรรมเผยแพร่การใช้งานโปรแกรม R และสิ่งประดิษฐ์ที่สร้างขึ้นในกิจกรรมสองกลุ่มที่กล่าวข้างต้น ในการวิเคราะห์ทางสถิติ การพัฒนาการใช้โปรแกรม R ให้ใช้ได้ง่ายขึ้น ประกอบด้วยการสร้างฟังก์ชันเพิ่มเติม โดยเฉพาะในด้านการจัดการข้อมูล และการแสดงผลการทำงานที่เข้าใจได้ง่าย ซึ่งเป็นจุดอ่อนที่สำคัญของ R และทำชุดฟังก์ชันเหล่านั้นเป็นแพคเกจเพื่อเผยแพร่แก่ผู้ต้องการใช้ และการพัฒนาส่วนติดต่อกับผู้ใช้ให้เป็นแบบรายการเมนูและหน้าต่างกราฟิกเช่นเดียวกับโปรแกรมทางการค้าอื่นๆ ในด้านการพัฒนาส่วนติดต่อกับผู้ใช้เป็นภาษาไทยนั้น ประกอบด้วยการทำความช่วยเหลือแบบ on line และแบบเป็นแฟ้มสำหรับฟังก์ชันและแพคเกจที่สร้างขึ้นนั้นเป็นภาษาไทย และการทำรายการเมนูและหน้าต่างกราฟิกต่างๆ ให้เป็นภาษาไทย ในด้านการจัดกิจกรรมเผยแพร่การใช้งานโปรแกรม R ได้จัดการประชุมเชิงปฏิบัติการให้แก่ผู้ใช้กลุ่มต่างๆ จัดการเรียนการสอนผ่านทางอินเตอร์เน็ต และการตีพิมพ์หนังสือเผยแพร่แก่ประชาชนและนักวิชาการ และยังต้องจัดทำเว็บไซต์เพื่อการเรียนรู้ แลกเปลี่ยนความคิดเห็นและการจัดกิจกรรมต่างๆ ที่เกี่ยวข้องกับ R

ผลการดำเนินงาน

ผลการดำเนินงานในกลุ่มกิจกรรมการพัฒนาการใช้โปรแกรม R ให้ใช้ได้ง่ายขึ้น ประกอบด้วยการแพคเกจชุดฟังก์ชันด้านการจัดการแฟ้มและข้อมูล ชื่อ ice มีฟังก์ชันที่คณะวิจัยได้สร้างขึ้น 50 ฟังก์ชัน แพคเกจ epid ซึ่งเป็นชุดฟังก์ชันพื้นฐานด้านระบาดวิทยา มีฟังก์ชันที่คณะวิจัยได้สร้างขึ้น 9 ฟังก์ชัน ความช่วยเหลือของฟังก์ชันที่สร้างขึ้นเหล่านี้มีทั้งที่เป็นภาษาอังกฤษและภาษาไทย และได้สร้างสภาพแวดล้อม ICE ซึ่งเป็นสภาพแวดล้อมที่เป็นรายการเมนูและหน้าต่างกราฟิก สามารถแสดงได้ทั้งภาษาไทยและภาษาอังกฤษ ประกอบด้วยแพคเกจหลักคือ ice.main และแพคเกจประกอบอื่นๆ ซึ่งเมื่อทำงานภายใต้โปรแกรม R จะเรียกเป็นสภาพแวดล้อม Integrated Computing Environment for R (R-ICE) ในการจัดกิจกรรมเพื่อเผยแพร่การใช้งาน

โปรแกรม R นั้น ได้จัดการประชุมเชิงปฏิบัติการวิเคราะห์ทางสถิติด้วยโปรแกรม R ขึ้น 13 ครั้ง ให้แก่ผู้สนใจทั้งภายในและภายนอกมหาวิทยาลัยสงขลานครินทร์ รวมถึงการจัดนอกสถานที่ด้วย ได้จัดทำเว็บไซต์ขึ้น 4 เว็บไซต์ มีเว็บไซต์สำหรับผู้สนใจโปรแกรม R โดยทั่วไป ซึ่งเป็นที่พบปะสนทนาของกลุ่มผู้ใช้ R เว็บไซต์สำหรับหลักสูตรทางอินเตอร์เน็ต เว็บไซต์สำหรับสภาพแวดล้อม R-ICE และเว็บบล็อกเผยแพร่การเขียนฟังก์ชันในโปรแกรม R และยังได้จัดทำ mirror ของ CRAN เพื่อให้ดาวน์โหลดโปรแกรมได้รวดเร็วยิ่งขึ้น ส่วนการตีพิมพ์หนังสือเพื่อส่งเสริมการใช้โปรแกรม R นั้น ได้ทำหนังสือขึ้นสองเล่ม คือ "การวิเคราะห์ทางสถิติด้วย R-ICE" ซึ่งมีกลุ่มเป้าหมายเป็นประชาชนทั่วไปที่สนใจการวิเคราะห์ข้อมูลทางสถิติ และ "คู่มือการใช้โปรแกรม R การใช้คำสั่งหรือฟังก์ชันพื้นฐาน" ซึ่งมีกลุ่มเป้าหมายเป็นนักวิชาการในสถาบันวิชาการต่างๆ ขณะนี้ทั้งสองเล่มกำลังอยู่ในระหว่างการดำเนินการจัดพิมพ์

วิจารณ์และสรุป

ในการดำเนินโครงการนี้ คณะผู้วิจัยได้เรียนรู้ในด้านเทคนิคต่างๆ มากมายในการเขียนฟังก์ชัน การทำแพคเกจ และการแปลภาษาตามวิธีการของโปรแกรม R และได้สร้างเครือข่ายกับนักพัฒนาโปรแกรม R ในระดับนานาชาติโดยการเข้าร่วมประชุม และนำเสนอผลงานเรื่องสภาพแวดล้อม R-ICE ที่พัฒนาขึ้นนี้ในการประชุม UseR! 2006 ที่กรุงเวียนนา ในการเผยแพร่การใช้งานโปรแกรม R นั้น การจัดประชุมเชิงปฏิบัติมีประสิทธิภาพในการเรียนรู้สูงสุด แต่ทำได้จำนวนจำกัด เนื่องจากต้องใช้ทรัพยากรมาก การจัดการเรียนการสอนทางอินเตอร์เน็ตไม่ได้ผลดีเท่าที่ควร เนื่องจากผู้เรียนอาจไม่อยู่ในสภาพที่พร้อมจะเรียน เช่นไม่สามารถติดตั้งโปรแกรมในเครื่องที่ใช้เรียนได้ และหากมีปัญหาเล็กน้อยก็มักจะเลิกล้มการเรียนได้ง่าย คิดว่าการทำหนังสือคู่มือน่าจะเผยแพร่ได้ในวงกว้างและผู้ใช้สามารถทำตามขั้นตอนและตัวอย่างในหนังสือได้ และไม่มีข้อจำกัดเรื่องเวลาและสถานที่ในการเรียนรู้ นอกจากนี้ การสร้างเว็บไซต์เพื่อเป็นที่รวมเสมือนของผู้ใช้ R ในประเทศไทยน่าจะนำไปสู่การสร้างเครือข่ายผู้ใช้ R ที่เป็นคนไทยได้อย่างต่อเนื่องไปในอนาคตระยะยาว

ทั้งนี้คณะผู้วิจัยได้ใช้โปรแกรม R ในการทำงานวิเคราะห์ทางสถิติ และการเรียนการสอนในหลักสูตรต่างๆ จึงจะยังคงทำงานนี้ต่อไปแม้จะสิ้นสุดโครงการวิจัยนี้แล้วก็ตาม โดยมีความคาดหวังว่าจะสามารถผลักดันให้มีการใช้โปรแกรม R ในประเทศไทย ทั้งนี้เพราะเรามั่นใจอย่างสูงว่า R เป็นซอฟต์แวร์วิเคราะห์ทางสถิติที่มีประสิทธิภาพสูง เป็นที่ยอมรับของนักสถิติผู้เชี่ยวชาญทั่วโลก และยังใช้และแจกจ่ายได้ฟรี โดยหวังว่าจะสามารถปลูกฝังเจตคติแก่สังคมไทยในการเคารพกฎหมายทรัพย์สินทางปัญญา และการไม่ใช้ซอฟต์แวร์ที่ได้มาโดยผิดกฎหมายอีกด้วย

Abstract

Nowadays, statistical analysis requires softwares that are capable of large dataset handling and sophisticated statistical methodologies, while most of them are high cost commercial softwares. **R**, a high capacity open source software for statistical analysis, is recently developed and rapidly accepted by world class statisticians and academic professions. The most distinguishing characteristic of **R** is its GPL (GNU General Public License) term where the software is open for free use, distribution, and modification. However, it is rarely known and used in Thailand. One of the important limitations is its strength as a computer language environment, thus, most of the documentations are available for programmers but not for statisticians and general users.

Objectives:

This project is aiming at development, distribution, and facilitation of using R as a free software for statistical analysis of research projects both in the fields of health science and statistical methodologies.

Methods and plans:

To achieve the above objectives, three groups of activities were planed: 1. modification of **R** for ease of use, 2. development of user interfaces in Thai, and 3. promotion of **R** and the products of the two previous activities for statistical analysis. Activities in development of **R** for ease of use included creation of new functions especially in data management and output display which were important weakness of **R**. Functions were bundled in a small set of packages that could be easily distributed to users. Graphical user interfaces (GUI) like those existing in other well known commercial statistical softwares was built. User interfaces in Thai included on-line and off-line help for those functions and packages and drop down menus and dialog windows in Thai. Activities for promotion of **R** in statistical analysis comprised workshops for different targeted users, internet-based courses, and web-sites for learning, discussion, and other activities related to **R**.

Results:

To modify **R** for ease of use, two packages are created. A package emphasizing on file and data management called "ice" comprises 50 newly created functions and "epid" package contains 9 basic epidemiological functions. Those functions have both English and Thai on-line and off-line help files. A GUI environment called "**ICE**" which has both English and Thai displays is also created. It contains one core package, "ice.main", and a set of associated packages. When it is running under **R**, Integrated Computing Environment for **R** (**R-ICE**) is the name for such an environment. To promote using **R** in statistical analysis, 13 workshops were held for university staff and students and those from other institutes. Four web sites have been set up for R users (UseR). One is a meeting place and discussion forum for Thai UseRs, one for internet-based courses, and one for **R-ICE** users. The last site is a web blog for **R** programming. A mirror site of CRAN is set up at a server in the Epidemiology Unit, Prince of Songkla University to facilitate rapid download of **R** software and packages within the university and the country. Two books are on the process of publishing. "Statistical Analysis with **R-ICE**" is targeted at general users

who are interested in statistical analysis. "Manual for Basic **R** Functions" is targeting at academic personnels.

Discussion and conclusion:

We have learned a lot of valuable **R** programming, package building, and internationalization techniques during the project. We have a network with **R** developers and **R** core team in UseR! 2006 conference in Vienna where we had an opportunity to present the **R-ICE** environment. In promoting the use of **R**, there are evidences that a workshop with computer practice is the most successful environment for learning and using **R** but it consumes a lot of resources while only a number of participates can be enrolled at a time. Internet-based courses is not a highly fruitful environment since participants may not be ready all the time of the courses and their problems may not be promptly solved and learning failure is ensured. Books and manuals can be widely distributed to target groups and they can follow the contents and practice examples at their own paces and places. Furthermore, web sites as virtual meeting places might lead to a last long network of Thai UseRs.

The research team is now using **R** in statistical analysis and teaching courses, thus, we still continue working on **R** even the project has come to an end. We hope we could continue promoting use of R in Thailand since we are best convinced that R is a free and highly efficient statistical software that is accepted by statistician communities worldwide and we also wish we could convince Thai communities an attitude of respecting intellectual property rights and not using any illegal softwares.