BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

ANGKANA HUANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2017

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

ANGKANA HUANG 5837702 ITCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE: APIRAK HOONLOR, Ph.D., PETER HADDAWY, Ph.D., DAMON W. ELLISON, Ph.D.

ABSTRACT

The abundance of bioinformatics tools has grown exponentially over the last three decades. Concurrently, many tools become outdated due to their discontinuation. Staying updated is highly difficult and is costly in terms of time and effort. The existing systems to ease tool discovery primarily attempts to index all the available tools according to their functionalities. Some of them provides the number of cites the tools received to indicate their popularity. The size of these lists have grown large calling for more targeted retrieval approaches. Since the tools are typically used in conjunction, a recent approach allows the users to browse the tools according to expert maintained pipelines. However, generating and updating these pipelines remains manual. Moreoever, the actual conjunct use of the tools are not provided. This thesis suggests an automated pipeline derivation method through literature mining and cross-citation analysis. The analysis patterns and the tool usage patterns were recovered from the data. Recommendation models were built from the data and the derived patterns. Through evaluating the models, actual tool selection behaviors were also understood. During 2009-2016, the tool functionalities with their popularity considered was highly predictive of whether they have been chosen. Along the period, substituting the overall popularity with the local popularity within the recovered analysis patterns became increasingly predictive and was on par in 2016. Such results implies that the recovered patterns resemble the pipelines used in community. Lastly, the pipelines were queried into the recommendation models to obtain the community accepted best-practices. These analysis patterns and best-practices can be used to inform experts regarding the status-quo of the field and can be used as guidelines for newcomers entering the field.

KEY WORDS : RECOMMENDATION SYSTEM / INFORMATION RETRIEVAL / BIOINFORMATICS

67 pages

การแนะนำเครื่องมือทางชีวสารสนเทศโดยอิงหลักฐานทางการใช้งาน

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

อังคณา หวัง 5837702 ITCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: อภิรักษ์ หุ่นหล่อ, Ph.D., Peter Haddawy, Ph.D., Damon W. Ellison, Ph.D.

บทคัดย่อ

จำนวนเครื่องมือทางชีวสารสนเทศได้ทวีขึ้นมากในสามทศวรรษที่ผ่านมา ในขณะเดียวกันเครื่องมือจำนวนมากก็ได้หยุดการพัฒนาลง การคงซึ่งความทันสมัยจึงเป็นไปโดยยาก ระบบในปัจจุบันมักมุ่งเน้นการทำดัชนีเครื่องมือและประโยชน์ใช้สอยของเครื่องมือนั้น จัดทำเป็นรายการเพื่อให้เอื้อต่อการค้นพบเครื่อง จำนวนเครื่องมือที่มีมากทำให้การแสดงรายการต้องมีความจำเพาะเจาะจงยิ่งขึ้น ด้วยเครื่องมือชีวสารสนเทศมักถูกใช้ร่วมกันอย่างเป็นลำดับขั้น เมื่อไม่นานมานี้ จึงมีระบบหนึ่งแสดงรายการโดยจัดหมวดหมู่เป็นลำดับตามลำดับขั้นการใช้งาน อย่างไรก็ดี ลำดับขั้นเหล่านี้ยังคงต้องถูกกำหนดและปรับเปลี่ยนโดยมนุษย์ การวิจัยนี้จึงมุ่งให้สามารถค้นพบลำดับขั้นการใช้งานที่มีอยู่ในวรรณกรรมโดยอัตโนมัติและลำดับขั้นที่ถูกค้นพบนั้นไปสร้างระบบแนะนำเครื่องมือชีวสารสนเทศ การประเมินประสิทธิภาพการแนะนำเครื่องมือได้ทำให้ทราบว่าในระหว่าง ค.ศ.2009-2016 ประโยชน์ใช้สอยของแต่ละเครื่องมือร่วมกับความนิยมโดยรวมของเครื่องมือนั้น ๆ ส่งผลสูงสุดต่อการเลือกเครื่องมือ ส่วนความนิยมของเครื่องมือที่จำเพาะต่อลำดับขั้นการใช้งานนั้นเริ่มมีอิทธิพลมากขึ้นในระยะหลังและสูงเทียบเท่ากับความนิยมโดยรวมในปีสุดท้ายของการศึกษา ผลการวิจัยนี้แสดงให้เห็นว่าลำดับขั้นการใช้งานที่ค้นพบโดยอัตโนมัตินี้มีความใกล้เคียงกับลำดับขั้นที่แท้จริงในข้อมูล ประโยชน์ใช้สอยที่อยู่ในลำดับขั้นเหล่านั้นถูกป้อนเข้าสู่ระบบแนะนำเครื่องมือที่สร้างขึ้นเพื่อแสดงชุดเครื่องมือที่เป็นที่ยอมรับและทันสมัยในปัจจุบัน ความรู้นี้เป็นประโยชน์ต่อการติดตามความเปลี่ยนแปลงในการใช้เครื่องมือชีวสารสนเทศ