ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do- Hạnh phúc

ĐỀ CƯƠNG ĐỀ TÀI LUẬN VĂN THẠC SỸ

1. **Tên đề tài:**

Tên tiếng Việt: **Xác định tập dữ liệu huấn luyện phù hợp nhằm hiệu chỉnh mô hình COCOMO**

Tên tiếng Anh: **Determining appropriate training sets for calibrating COCOMO**

2. **Ngành và mã ngành** đào tạo**:** Ngành Công nghệ thông tin, mã ngành**:**

3. **Họ và tên học viên** thực hiện đề tài**:** Huỳnh Thị Phương Thủy
   Địa chỉ email, điện thoại liên lạc: thuyhuynh409@yahoo.com , 0918 340 741

4. **Người hướng dẫn:** TS. Nguyễn Văn Vũ
   Địa chỉ email, điện thoại liên lạc : nvu@fit.hcmus.edu.vn , 0908 175 957

5. **Tổng quan tình hình NC:**

5.1. Giới thiệu chung

Ngành công nghệ phần mềm đã và đang phát triển mạnh, cùng với các công cụ hỗ trợ và kỹ thuật tiên tiến hơn, các mô hình phát triển phần mềm mới cũng liên tục ra đời, kế thừa và phát triển hơn mô hình cũ. Cũng vì thế mà năng suất phần mềm ngày càng được nâng cao hơn. Việc ước lượng chi phí và thời gian phát triển phần mềm đòi hỏi phải có sự điều chỉnh lại để cải thiện độ chính xác.

Ước lượng hay dự đoán trước một vấn đề nào đó luôn luôn cần thiết và quan trọng. Trong kinh doanh, ước lượng để từ đó có những quyết định đầu tư phù hợp về tiền bạc, thời gian, con người, v.v… Tuy nhiên, nếu ước lượng quá cao hay quá thấp

hơn mức thực tế đều có thể làm thiệt hại như không có được hợp đồng khi đấu thầu, hoặc có được hợp đồng nhưng bị thua lỗ về tài chính.

COCOMO là một trong những mô hình ước lượng chi phí và thời gian phát triển phần mềm. Mô hình này do Boehm là người đầu tiên xây dựng vào năm 1981, kế đó là phiên bản COCOMO II vào năm 2000 và hiện đang được tiếp tục mở rộng và phát triển một phiên bản COCOMO III. COCOMO đã và đang thu hút sự chú ý của cộng đồng nghiên cứu và được rất nhiều công ty áp dụng trong thực tế. Mặc dù vậy, cải thiện mức độ chính xác vẫn luôn là bài toán quan trọng đặt ra cho những người nghiên cứu và là mong đợi của các nhà phát triển trong ngành công nghiệp phần mềm.

Trong đề tài này, tôi sẽ nghiên cứu về mức độ ảnh hưởng từ các bộ dữ liệu khác nhau được rút trích theo thời gian và theo miền dự án lên độ chính xác của mô hình COCOMO. Cụ thể, tôi sẽ đánh giá độ chính xác của mô hình COCOMO được hiệu chỉnh dựa trên bộ dữ liệu huấn luyện được rút trích thông qua cửa sổ di động (moving window) [2] và bộ dữ liệu theo miền dự án.

5.2. Các thách thức

- Ước lượng hay dự đoán trước một điều gì bao giờ cũng là một bài toán khó, bởi mức độ chính xác phụ thuộc vào rất nhiều yếu tố. Ước lượng phần mềm cũng vậy, *dựa vào bộ dữ liệu* chỉ là một trong số các nghiên cứu làm ảnh hưởng đến độ chính xác của ước lượng.

- Càng khó khăn hơn khi bộ dữ liệu là các mẫu không ngẫu nhiên được thu thập từ các tổ chức và lĩnh vực kinh doanh cụ thể. Hơn nữa, đa phần các tổ chức công ty thường không muốn chia sẻ dữ liệu của mình với cộng đồng công nghệ phần mềm

- Từ bộ dữ liệu có sẵn, việc lựa chọn ra các tập dữ liệu con thích hợp cũng không phải dễ. Các thuật toán khác nhau có thể mang lại các kết quả khác nhau. Cũng có thể xảy ra là các tập con khác nhau có kích thước bằng nhau lại cho cùng kết quả**.**

5.3. Tình hình nghiên cứu

- ***Tình hình nghiên cứu ngoài nước***

Nghiên cứu về ước lượng chi phí phần mềm đã có một lịch sử khá dài [3]. Tuy nhiên, rất ít công trình tập trung nghiên cứu đến ảnh hưởng của miền dữ liệu lên chi phí phần mềm.

Các mô hình ước lượng chi phí phát triển phần mềm thường được xây dựng và được đánh giá từ các dự án trong quá khứ. Câu hỏi quan trọng được đặt ra là chọn những dự án nào dùng làm bộ dữ liệu huấn luyện để xây dựng mô hình. Nên chăng là dùng toàn bộ các dự án trong quá khứ hay là nên chọn một tập con các dự án nào đó? Khoảng cách thời gian hay tuổi của dự án chính là một trong các yếu tố để chọn lựa dự án nào được dùng cho bộ dữ liệu huấn luyện để hiệu chỉnh mô hình ước lượng.

Công trình S1 [5], Lokan và Mandes đã nghiên cứu về việc dùng moving windows trên bộ dữ liệu 228 dự án của các tổ chức độc lập, được lấy từ kho dữ liệu ISBSG [7]. Bộ dữ liệu huấn luyện được xác định bởi N dự án đã hoàn thành gần đây nhất. Họ nhận thấy N (window size) có giá trị lớn sẽ ảnh hưởng tốt hơn so với N có giá trị nhỏ. Tốt nhất N nên trên dưới 75 dự án.

Công trình S2[1], Amasaki nghiên cứu dùng moving window với kích thước khác nhau trên bộ dữ liệu được lấy từ kho dữ liệu PROMISE. Kỹ thuật ước lượng dựa vào analogy. Họ cũng thấy rằng việc dùng moving windows có cải thiện độ chính xác của ước lượng mặc dù không nhiều.

Công trình S3 [2], Amasaki và Lokan đã nghiên cứu dùng moving window trên bộ dữ liệu của S1 với cả hai kỹ thuật ước lượng truy hồi tuyến tính (Liner Regression - LR) và ước lượng dựa theo tính tương đồng (Estimation by Analogy - EbA). Kết quả là có sự khác nhau về mức độ chính xác của ước lượng khi có áp dụng và không có áp dụng moving windows đối với cả hai kỹ thuật ước lượng LR và EbA. Tuy nhiên mức độ ảnh hưởng đối với EbA ít hơn so với LR.

Công trình S4 [6], Lokan và Mendes tiếp tục nghiên cứu trên bộ dữ liệu của S1 và S3 để nghiên cứu ảnh hưởng của miền window size lên độ chính xác của ước lượng. Họ chỉ ra rằng việc sử dụng các tập dữ liệu theo thời gian có làm ảnh hưởng đến độ chính xác của ước lượng, nhưng mức độ thấp hơn so với các tập dữ liệu dựa trên số lượng dự án.

Lefley và Shepperd [4], và Sentas [10] đã dựa vào yếu tố thời gian là cơ bản khi phân chia bộ dữ liệu thành các tập dữ liệu huấn luyện và tập dữ liệu kiểm tra để so sánh các mô hình ước lượng phần mềm. Lokan và Mendes cũng so sánh các mô hình mà tập dữ liệu huấn luyện và tập dữ liệu kiểm tra khi được chọn một cách ngẫu nhiên và khi được chọn dựa trên yếu tố thời gian. Tuy nhiên họ vẫn chưa tìm được sự ảnh hưởng đáng kể lên độ chính xác của ước lượng.

MacDonell và Shepperd [8] cũng dùng moving windows trong nghiên cứu làm cho dữ liệu ở các giai đoạn đầu của một dự án tốt hơn và có thể dùng nó để ước lượng cho các giai đoạn sau. Họ nhận thấy độ chính xác được cải thiện hơn khi dùng moving window với kích thước là 5 dự án gần đây nhất, được dùng làm dữ liệu huấn luyện hơn là dùng tất cả các dự án hoàn thành.

- ***Tình hình nghiên cứu trong nước***

Ở Việt Nam đã có một số đề tài, luận văn nghiên cứu về các mô hình ước lượng chi phí phần mềm, đa phần tập trung nghiên cứu về tổng quan, xây dựng mô hình và kỹ thuật thực hiện. Trong hiểu biết hạn hẹp của tôi thì nghiên cứu liên quan đến độ chính xác của mô hình COCOMO được tìm thấy trên bài báo khoa học [9] công bố năm 2011 đăng trên kỷ yếu của Hội nghị quốc tế lần thứ 7 về các mô hình ước lượng phần mềm, được nghiên cứu bởi TS.Nguyễn Văn Vũ là người Việt Nam cùng 2 đồng sự là người nước ngoài, ông LiGuo Huang và ông Barry Boehm. Sau đó, đầu năm 2015, đề tài chính thức được Quỹ Phát triển Khoa học và Công nghệ Quốc gia tài trợ thực hiện do TS.Nguyễn Văn Vũ làm chủ nhiệm . Mục tiêu nghiên cứu nhằm cải tiến COCOMO dựa trên Phân tích Xu hướng của năng suất dự án

và các nhân tố ảnh hưởng theo thời gian, trong đó có phân tích ảnh hưởng của các nhân tố khác, bao gồm lĩnh vực kinh doanh và quốc gia (nghĩa là phân tích theo miền dự án), lên năng suất và xu hướng năng suất.

Như vậy, hiện vẫn chưa có đề tài nào tập trung nghiên cứu về việc làm tăng mức độ chính xác của mô hình COCOMO bằng cách phân chia tập dữ liệu vốn có thành các tập con phù hợp hơn theo phương pháp moving windows và theo miền dự án.

6. Tính khoa học và tính mới của đề tài:

Như đã trình bày về tình hình nghiên cứu trong và ngoài nước, các nghiên cứu trên đều cho thấy việc dùng moving window để chọn các tập dữ liệu khác nhau từ các nguồn khác nhau có thể ảnh hưởng lên độ chính xác của ước lượng. Tuy nhiên vẫn chưa có nghiên cứu về việc dùng moving window trên bộ dữ liệu COCOMO để đánh giá mức độ ảnh hưởng của miền dự án lên độ chính xác của mô hình COCOMO vốn là mô hình ước lượng chi phí phổ biến nhất trong lĩnh vực công nghệ phần mềm.

COCOMO 81 là mô hình ước lượng đầu tiên do Boehm xây dựng để ước tính chi phí và thời gian phát triển phần mềm dựa trên 17 yếu tố ảnh hưởng. Sau đó, mô hình COCOMO cần thiết phải được nâng cấp để thích ứng với những thay đổi lớn trong ngành công nghệ phần mềm, Dr. Boehm và nhóm nghiên cứu của ông tại Đại học Nam California (University of Southern California) đã phát triển, nâng cấp mô hình này lên COCOMO II bằng cách bổ sung thêm 5 yếu tố chi phí nữa để phản ánh tốt hơn việc phát triển phần mềm hiện tại. 22 yếu tố quan trọng nhất trong COCOMO II có mức độ ảnh hưởng khác nhau đến chi phí và thời gian phát triển phần mềm, được rút ra từ việc phân tích dữ liệu của 161 dự án đã hoàn thành từ năm 1970 đến năm 2009.

Đề tài này tôi sẽ nghiên cứu về việc áp dụng moving window để phân chia 341 dự án của tập dữ liệu COCOMO và tập dữ liệu bổ sung đã thu thập được trong suốt những năm 1970 đến 2009 thành các tập con phù hợp hơn (theo thời gian hoàn thành, theo số lượng dự án,..).

Bộ dữ liệu 341 dự án mà chúng tôi sử dụng trong đề tài này được cung cấp bởi 25 tổ chức từ 4 quốc gia bao gồm Mỹ, Brazil, Thái Lan và Việt Nam. (có14 dự án từ các công ty phần mềm tại Việt Nam)

Kết quả của đề tài sẽ đưa ra những bằng chứng thực nghiệm cho việc xác định kích thước window size, tạo ra các tập con phù hợp nhất dùng làm bộ dữ liệu huấn luyện và kiểm tra. Các mô hình thu được sẽ phản ảnh tốt hơn thực tiễn phát triển phần mềm nói chung và Việt Nam nói riêng, nhờ đó sẽ cung cấp các ước lượng chính xác hơn mô hình COCOMO tổng quát.

7. Mục tiêu, đối tượng và phạm vi nghiên cứu

7.1. Mục tiêu

Nghiên cứu khả năng tăng độ chính xác của mô hình COCOMO thông qua việc áp dụng phương pháp moving window nhằm chọn tập dữ liệu huấn luyện thích hợp. Mục tiêu này sẽ trả lời hai câu hỏi nghiên cứu:

RQ1: Độ chính xác của mô hình COCOMO có thể được cải thiện hay không nếu dữ liệu huấn luyện được rút trích từ *K* dự án hoàn thành trước đó?

RQ2: Độ chính xác của mô hình COCOMO có thể được cải thiện hay không nếu dữ liệu huấn luyện được rút trích từ những dự án hoàn thành trong khoảng *N* năm trước đó?

- Tìm hiểu mức độ ảnh hưởng của miền dự án lên độ chính xác của mô hình COCOMO. Cụ thể, câu hỏi nghiên cứu RQ3 sẽ được đánh giá:

  RQ3: Độ chính xác của mô hình COCOMO có thể được cải thiện hay không nếu nhân tố miền dự án được sử dụng?

- Đưa ra các đề xuất nhằm cải thiện độ chính xác của mô hình COCOMO, cụ thể:
  - Định nghĩa và xác định được kích thước moving windows.

- Áp dụng moving window tạo ra các tập dữ liệu con của tập dữ liệu COCOMO.
- Xây dựng mô hình ước lượng mới dựa vào mỗi tập dữ liệu con
- Đánh giá mô hình mới thông qua chỉ số MRE và PRED
- Rút ra các kết luận về độ chính xác của mô hình lên các miền dữ án khác nhau.

7.2. Đối tượng và phạm vi áp dụng:

- Công thức ước lượng công sức COCOMO II
- Bộ dữ liệu COCOMO gồm 341 dự án

8. Nội dung, phương pháp dự định nghiên cứu:

8.1. Nội dung 1: Xử lý dữ liệu theo phương pháp moving window

- **Mục tiêu:** Chia bộ dữ liệu COCOMO gồm 341 dự án thành các tập dữ liệu con (dựa theo phương pháp moving window) để huấn luyện hay hiệu chỉnh mô hình
- **Phương pháp:** Xác định kích thước cửa sổ theo số dự án và khoảng thời gian. Mỗi lần di chuyển cửa sổ sẽ xác định được một tập dữ liệu con

8.2. Nội dung 2: Xử lý dữ liệu theo miền dự án

- **Mục tiêu:** Chia bộ dữ liệu COCOMO gồm 341 dự án thành các tập dữ liệu con dựa trên miền dự án để huấn luyện hay hiệu chỉnh mô hình
- **Phương pháp:** Sử dụng thông tin miền dự án để xác định các tập dữ liệu có cùng miền [9].

8.3. Nội dung 3: Hiệu chỉnh COCOMO trên mỗi tập dữ liệu con

- **Mục tiêu:** Làm tăng độ chính xác cho mô hình ước lượng COCOMO trên mỗi tập dữ liệu con được phân chia trong Nội dung 1 và 2 ở trên.
- **Phương pháp:**

Từ mỗi tập dữ liệu con, hiệu chỉnh tham số A và B trong công thức COCOMO II bằng phương pháp hồi quy tuyến tính nhằm tạo ra mô hình ước lượng mới.

$$\mathbf{PM = A \times Size^{E} \times \prod_{i=1}^{17} EM_i}$$

$$\mathbf{với\ E = B + 0.01 \times \sum_{j=1}^{5} SF_j}$$

8.4. Nội dung 4: Đánh giá độ chính xác của mô hình COCOMO trên từng tập dữ liệu con

- **Mục tiêu:** Đánh giá độ chính xác của mô hình ước lượng dựa trên phương pháp moving window và miền dự án
- **Phương pháp:** Xác định các độ đo MRE (magnitude of relative errors) và PRED (preduction level):

Mức độ lỗi (MRE)

$$\mathbf{MRE} = \frac{|\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{o}|}{\mathbf{y_i}}$$

Trung bình của MRE (MMRE)

$$\mathbf{MMRE} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{MRE}$$

Độ chính xác: *PRED(l) = k/N*

(k là số các ước lượng với giá trị *MRE ≤ l)*

Lặp lại cho tới khi tất cả các tập con được tính MRE và PRED

Kết luận: so sánh độ chính xác của mô hình COCOMO đối với các window durations khác nhau (các miền dự án khác nhau)

9. Kế hoạch thực hiện nghiên cứu:

Nội dung	Thời gian
• Tìm hiểu tổng quan về mô hình ước lượng công sức COCOMO	3/2015

• *Xin ý kiến của Thầy Vũ*	
• Xử lý Bộ dữ liệu COCOMO gồm 341 dự án • *Xin ý kiến của Thầy Vũ*	4/2015
• Khảo sát độ chính xác của mô hình COCOMO đối với các miền dự án khác nhau (các window duration khác nhau) • *Xin ý kiến của Thầy Vũ*	5/2015
• Viết luận văn	6/2015
– *Nộp luận văn cho Thầy Vũ phiên bản 1*	*7/2015*
– *Hiệu chỉnh luận văn* – *Nộp cho Thầy Vũ phiên bản 2,3,..*	*7/2015*
• Nộp luận văn cho Trường	**8/2015**
• Bảo vệ luận văn	**9/2015**

10. Tài liệu tham khảo

[1] Amasaki, S., Takahara Y., Yokogawa, T. (2011), "Performance evaluation of windowing approach on effort estimation by analogy", *IWSM/Mensura'11*, 188–195.

[2] Amasaki, S., Lokan, C. (2012), "The effects of moving windows to software estimation: comparative study on linear regression and estimation by analogy", *IWSM/Mensura'12*.

[3] Jorgensen, M., Shepperd, M.J. (2007), "A systematic review of software development cost estimation studies", IEEE Trans. Software Eng. 33 (1) 33–53.

[4] Lefley, M., Shepperd, M.J. (2003), "Using genetic programming to improve software effort estimation based on general data sets", *GECCO*, 2477–2487.

[5] Lokan, C., Mendes, E. (2009), "Applying moving windows to software effort estimation", *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society*, 111–122.

[6] Lokan, C., Mendes, E. (2012), "Investigating the use of duration-based moving windows to improve software effort prediction", *K. R. P. H. Leung, P. Muenchaisri (Eds.), APSEC, IEEE*, 818–827.

[7] Lokan, C. and Mendes E. (2014), "Investigating the use of duration-based moving windows to improve software effort prediction: A replicated study", *Information and Software Technology* .

[8] MacDonell, S.G., Shepperd, M. (2010), "Data accumulation and software effort prediction", *Proceedings of the 2010 ACM IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, New York, NY, USA*, 31:1–31:4.

[9] Nguyen, V., L. Huang, L. and Boehm, B. (2011), "An analysis of trends in productivity and cost drivers over years", *Proceedings of the 7th International Conference on Predictive Models in Software Engineering*.

[10] Sentas, P., Angelis, L., Stamelos, I., Bleris, G.L. (2005), "Software productivity and effort prediction with ordinal regression", *Information & Software Technology* 47(1), 17–29.

TP.HCM, ngày tháng năm 2015

NGƯỜI HƯỚNG DẪN	**HỌC VIÊN**
(Họ tên và chữ ký)	*(Họ tên và chữ ký)*
TS. Nguyễn Văn Vũ	**Huỳnh Thị Phương Thủy**