Phụ lục 3

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

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

THÔNG TIN VỀ LUẬN ÁN TIẾN SĨ

1. Họ và tên nghiên cứu sinh: Phạm Thị Ngân ..... 2. Giới tính: Nữ.....

3. Ngày sinh: 18/07/1983 ..... 4. Nơi sinh: Quảng Ninh .....

5. Quyết định công nhận NCS số 858/QĐ-ĐT ngày 29/10 /2012 của Hiệu trưởng trường Đại học Công nghệ

6. Các thay đổi trong quá trình đào tạo: .....

(*ghi các hình thức thay đổi và thời gian tương ứng*)

7. Tên đề tài luận án: Nghiên cứu cải tiến phân lớp đa nhãn và ứng dụng.....

(*tên luận án chính thức đề nghị bảo vệ cấp Đại học Quốc gia*)

8. Chuyên ngành: Hệ thống thông tin..... 9. Mã số: 62.48.05.01 .....

10. Cán bộ hướng dẫn khoa học: PGS.TS Hà Quang Thụy, PGS.TS Phan Xuân Hiếu ...

(*ghi rõ chức danh khoa học, học vị, họ và tên)*

11. Tóm tắt các **kết quả mới** của luận án:

- Đề xuất một thuật toán phân lớp đa nhãn khai thác tính riêng biệt dựa trên phân cụm bán giám sát (Thuật toán MASS [PTNgan5]) trên cơ sở áp dụng một chiến lược tham lam khi tích hợp hai thuật toán LIFT [Zhang15a] và TESC [Zhang15b]. Mối liên quan giữa các nhãn lớp được quy tụ vào các cụm dữ liệu mẫu có cùng tập nhãn. Tính hiệu quả của thuật toán đề xuất đã được luận giải và kiểm chứng bằng thực nghiệm.
- Đề nghị hai mô hình biểu diễn dữ liệu cho phân lớp đa nhãn là mô hình biểu diễn dữ liệu đồ thị khoảng cách được đề xuất trong [PTNgan4] nhằm khai thác các thông tin bậc cao về trật tự và khoảng cách đặc trưng trong văn bản và mô hình biểu diễn dữ liệu chủ đề ẩn được đề xuất trong [PTNgan3] nhằm khai thác các thông tin ngữ nghĩa ẩn trong văn bản để làm giàu thêm các đặc trưng cho mô hình. Tính hiệu quả của các mô hình biểu diễn văn bản đã được luận giải và kiểm chứng bằng thực nghiệm.
- Đề xuất một mô hình phân lớp đa nhãn MULTICS trong [PTNgan6] dựa trên khai thác thuật toán phân lớp đa nhãn bán giám sát trong [PTNgan5] kết hợp với tiếp cận biểu diễn dữ liệu chủ đề ẩn khai thác thông tin ngữ nghĩa ẩn trong văn bản làm giàu đặc trưng cho thuật toán phân lớp.

- Luận án đóng góp vào dòng nghiên cứu trên thế giới và trong nước về học máy đa nhãn trong văn bản tiếng Việt thông qua công bố 06 bài báo khoa học tại các ấn phẩm khoa học quốc tế có uy tín.
- Các kết quả nghiên cứu từ luận án đã được thực nghiệm, đánh giá kết quả cho thấy có tiềm năng ứng dụng thực tế.

12. Khả năng ứng dụng trong thực tiễn: (*nếu có*) Ứng dụng trong các bài toán liên quan đến phân lớp văn bản nói chung

13. Những hướng nghiên cứu tiếp theo: Trong thời gian tiếp theo, nghiên cứu sinh sẽ tiếp tục nghiên cứu các hướng giải quyết cho các hạn chế còn tồn tại của luận án và tiếp tục triển khai các đề xuất để hoàn thiện hơn các giải pháp cho phân lớp đa nhãn.

- Một là, thuật toán MULTICSLearn cần được phân tích sâu sắc hơn đặc biệt ở khía cạnh độ phức tạp thời gian tính toán trong một vùng hoặc toàn bộ miền ứng dụng. Cần đưa ra các phân tích đánh giả để luận giải được tính hiệu quả của chiến thuật tham lam được dùng trong thuật toán ít nhất về độ phức tạp thời gian trong trường hợp xấu nhất.
- Hai là, các kỹ thuật giảm chiều dữ liệu tiên tiến cho phân lớp đa nhãn cần được nghiên cứu để áp dụng sáng tạo vào các bài toán ứng dụng trong luận án.
- Ba là, khảo sát miền ứng dụng dữ liệu ảnh, nghiên cứu các mô hình và giải pháp phân lớp đa nhãn – đa thể hiện đối với dữ liệu ảnh nhằm làm phù hợp với quá trình tiến hóa của phân lớp dữ liệu.

14. Các công trình đã công bố có liên quan đến luận án: .....

- Thi-Ngan Pham, Le-Minh Nguyen, Quang-Thuy Ha (2012). *Named Entity Recognition for Vietnamese documents using semi-supervised learning method of CRFs with Generalized Expectation Criteria*. IALP 2012: 85-89
- Thi-Ngan Pham, Thi-Thom Phan, Phuoc-Thao Nguyen, Quang-Thuy Ha (2013). *Hidden Topic Models for Multi-label Review Classification: An Experimental Study*. Computational Collective Intelligence. Technologies and Applications, Lecture Notes in Computer Science Volume 8083:603-611.
- Thi-Ngan Pham, Thi-Hong Vuong, Thi-Hoai Thai, Mai-Vu Tran, Quang-Thuy Ha (2016). *Sentiment Analysis and User Similarity for Social Recommender System: An Experimental Study*. Lecture Notes in Electrical Engineering (376): 1147-1156
- Thi-Ngan Pham, Van-Hien Tran, Tri-Thanh Nguyen, Quang-Thuy Ha (2017). *Exploiting Distance graph and Hidden Topic Models for Multi-label Text Classification*. ACIIDS 2017. Studies in Computational Intelligence, Volume 710 (Advanced Topics in Intelligent Information and Database Systems): 321-331.
- Thi-Ngan Pham, Van-Quang Nguyen, Duc-Trong Dinh, Tri-Thanh Nguyen, Quang-Thuy Ha (2017). *MASS: a Semi-supervised Multi-label Classification Algorithm With specific Features*. ACIIDS 2017. Studies in Computational

Intelligence, Volume 710 (Advanced Topics in Intelligent Information and Database Systems): 37-47.

- Thi-Ngan Pham, Van-Quang Nguyen, Van-Hien Tran, Tri-Thanh Nguyen, and Quang-Thuy Ha (2017). *A semi-supervised multi-label classification framework with feature reduction and enrichment*. Journal of Information and Telecommunication, 1-14.

Ngày tháng năm 20	*Ngày tháng năm 20*
Xác nhận của cán bộ hướng dẫn	**Nghiên cứu sinh**
(Kí và ghi rõ họ tên)	*(Kí và ghi rõ họ tên)*

VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name : Pham Thi Ngan ..... 2. Sex: Female .....

3. Date of birth: 18/07/1983 ..... 4. Place of birth: Quang Ninh.....

5. Admission decision number: ..858/QĐ-ĐT .. Dated 29/10/2012 .....

6. Changes in academic process: ....

(List the forms of change and corresponding times)

7. Official thesis title: Study on improving multi-label classification and applications ...

8. Major: Information System ..... 9. 62.48.05.01 .....

10. Supervisors: Assoc. Prof. Ha Quang Thuy, Assoc. Prof. Phan Xuan Hieu.....

(Full name, academic title and degree)

11. Summary of the **new findings** of the thesis:

- I proposed a MLC algorithm (MASS algorithm in [PTNgan5]) which exploits specific features from semi-clustering technique based on applying greedy procedure by combining two algorithms of LIFT[Zhang15a] and TESC [Zhang15b]. The efficiency of the proposed algorithm is proved by experiments.
- I proposed two models of data representation for MLC. The former is the model of data representation based on distance graph [PTNgan4] exploiting high level information of order and distance between features in document. The latter is the model of data representation based on hidden topic model LDA [PTNgan3] which exploits hidden semantic meanings in document for enriching features. The efficiency of the proposed algorithm is proved by experiments.
- I proposed an model of multi-label classification MULTICS in [PTNgan6] which applies semi-supervised multi-label classification algorithm in [PTNgan] in combination with method of data representation based on hidden topic model LDA to improve the performance of the multi-label classifier.
- Thesis contributes to national and international research community about multi-label classification on Vietnamese by publishing 06 science papers on prestigious international publishers.

12. Practical applicability, if any: Applying in problems related text classification .....

13. Further research directions, if any: I focus on solving the shortcomings of the thesis and improving the proposals for multi-label classification. .....

- Firstly, focosing on analysing the semi-supervised multi-label classification, especially on more details about the algorithm complexity on the data and the greedy approach in choosing the dominant label.
- Secondly, studying and applying new dimensional extraction methods in the proposed models.
- Finally, studying and applying the proposed models in the domain of image; studying and proposing models for multi-label multi-instance classification.

14. Thesis-related publications: .....

- Thi-Ngan Pham, Le-Minh Nguyen, Quang-Thuy Ha (2012). *Named Entity Recognition for Vietnamese documents using semi-supervised learning method of CRFs with Generalized Expectation Criteria*. IALP 2012: 85-89
- Thi-Ngan Pham, Thi-Thom Phan, Phuoc-Thao Nguyen, Quang-Thuy Ha (2013). *Hidden Topic Models for Multi-label Review Classification: An Experimental Study*. Computational Collective Intelligence. Technologies and Applications, Lecture Notes in Computer Science Volume 8083:603-611.
- Thi-Ngan Pham, Thi-Hong Vuong, Thi-Hoai Thai, Mai-Vu Tran, Quang-Thuy Ha (2016). *Sentiment Analysis and User Similarity for Social Recommender System: An Experimental Study*. Lecture Notes in Electrical Engineering (376): 1147-1156
- Thi-Ngan Pham, Van-Hien Tran, Tri-Thanh Nguyen, Quang-Thuy Ha (2017). *Exploiting Distance graph and Hidden Topic Models for Multi-label Text Classification*. ACIIDS 2017. Studies in Computational Intelligence, Volume 710 (Advanced Topics in Intelligent Information and Database Systems): 321-331.
- Thi-Ngan Pham, Van-Quang Nguyen, Duc-Trong Dinh, Tri-Thanh Nguyen, Quang-Thuy Ha (2017). *MASS: a Semi-supervised Multi-label Classification Algorithm With specific Features*. ACIIDS 2017. Studies in Computational Intelligence, Volume 710 (Advanced Topics in Intelligent Information and Database Systems): 37-47.
- Thi-Ngan Pham, Van-Quang Nguyen, Van-Hien Tran, Tri-Thanh Nguyen, and Quang-Thuy Ha (2017). *A semi-supervised multi-label classification framework with feature reduction and enrichment*. Journal of Information and Telecommunication, 1-14.

Date: ……..
Signature: ……
Full name: ……

Date: ……..
Signature: ……
Full name: ……