# ỨNG DỤNG NGÔN NGỮ LẬP TRÌNH VBA TRONG EXCEL XÂY DỰNG CÔNG CỤ TỔNG HỢP CÂU HỎI MỞ VÀ KHAI THÁC DỮ LIỆU DẠNG CHỮ

*Nguyễn Thế Hưng**

***Tóm tắt:***

*Tổng hợp, xử lý thông tin của câu hỏi mở và khai phá dữ liệu dạng chữ luôn là vấn đề khó đối với mọi nghiên cứu. Bài viết này bàn về việc sử dụng ngôn ngữ lập trình VBA trong Excel để xây dựng công cụ tổng hợp mới thuận tiện nhiều đối tượng sử dụng khác nhau.*

Câu hỏi mở luôn là một phần quan trọng trong các nghiên cứu nói chung và trong bảng hỏi nói riêng. Câu hỏi mở rất hữu ích khi thu thập thông tin trong các nghiên cứu nhằm thăm dò hành vi, đánh giá của người trả lời về một chủ đề cụ thể nhất là đối với các biến không so sánh được như các nhận xét, đánh giá về sản phẩm hoặc thói quen tiêu dùng. Ngoài ra, câu hỏi mở còn đặc biệt hữu hiệu giúp cho việc thu thập thông tin được chính xác và đầy đủ hơn khi người thiết kế bảng hỏi không tiên lượng hết các phương án trả lời khi xây dựng bảng hỏi. Câu hỏi mở khiến người được phỏng vấn cảm thấy thỏa mái, không bị gò bó trong các câu hỏi đóng do vậy sẽ thu thập được các thông tin đa chiều của nhiều đối tượng trả lời khác nhau. Trong nhiều trường hợp, câu hỏi mở cũng được dùng để kiểm tra xem người trả lời có hiểu đúng ý trong câu hỏi đóng hay không thông qua các câu hỏi dùng để kiểm tra như: Tại sao?

Tuy nhiên, trong việc xử lý, tổng hợp dữ liệu câu hỏi mở thường bị lãng quên hoặc quan tâm không đúng mức không phải vì nó không quan trọng mà do khó tổng hợp và xử lý thông tin. Việc tổng hợp, xử lý thông tin từ câu hỏi mở hiện nay thường được thực hiện thông qua hai cách: (1) Đọc từng câu trả lời rồi tóm lược ý để phân tích; hoặc (2) Dùng các phần mềm chuyên dụng như Atlas.ti hoặc nViVo để tổng hợp. Tuy nhiên, cả hai cách trên đều có những hạn chế nhất định.

Cách thứ nhất tuy không tốn chi phí tài chính trực tiếp nhưng cần nhiều công sức để tổng hợp, dễ dẫn đến những sai số phi mẫu trong quá trình tổng hợp. Với các cuộc khảo sát quy mô lớn, số lượng câu trả lời nhiều, việc tổng hợp bằng cách "thủ công" này sẽ cần rất nhiều thời gian và kết quả thu được khó kiểm soát thậm chí không hoàn toàn chính xác.

Cách thứ hai đòi hỏi cần có nguồn tài chính tương đối để mua bản quyền sử dụng các phần mềm từ nước ngoài. Tuy nhiên, những chương trình này chưa thân thiện với người dùng và chúng chưa xử lý tốt đối với các định dạng phông chữ tiếng Việt khác nhau.

Để khắc phục các hạn chế trên, tác giả sử dụng ngôn ngữ lập trình VBA (Visual Basic for Application) trong Excel xây dựng công cụ OQA để khai phá dữ liệu dạng chữ, ứng dụng để xử lý, tổng hợp câu hỏi mở.

* *Vụ Thống kê Giá*

Công cụ này được thiết kế và sử dụng trên nền tảng của Microsoft Excel nên thân thiện với người dùng, thuận tiện sử dụng với mọi đối tượng sử dụng khác nhau. Ngoài ra, công cụ này không yêu cầu cài đặt thêm ứng dụng nên người dùng không cần phải cặt đặt bổ sung bất cứ chương trình nào vào máy tính, do vậy đáp ứng được nhiều đối tượng sử dụng khác nhau.

Khi đã có các câu trả lời, người dùng chuyển các câu trả lời của câu hỏi mở vào 1 sheet trong công cụ OQA để tổng hợp thông qua 2 module xử lý như sau:

*Module1: Tổng hợp câu hỏi mở theo các từ khóa định sẵn*

Khi xây dựng bảng hỏi, người thiết kế câu hỏi mở cũng đã hướng đến một số từ khóa nhất định mang những thông tin cần thu thập. Những từ khóa có thể là một từ, một cụm hoặc một số cụm từ cụ thể (gọi chung là từ khóa).

Ví dụ: Khảo sát về kiến nghị của người lao động có những từ hoặc cụm từ như: "tăng lương", "tăng lương" và "giảm giờ làm" là những từ khóa.

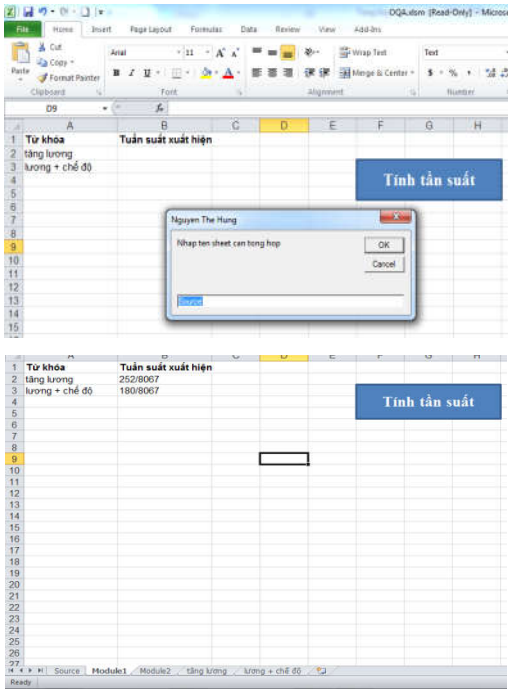
Để tổng hợp những câu trả lời có chứa các cụm từ riêng rẽ như trên, người dùng mở cộng cụ OQA và gõ từng từ/ cụm từ vào từng dòng của cột Từ khóa trong sheet Module1, ví dụ: Tăng lương, giảm giờ làm.

Hoặc muốn tổng hợp các câu trả lời có chứa đồng thời nhiều cụm từ cần nghiên cứu, người dùng gõ các cụm từ này vào từng dòng cột Từ khóa trong sheet Module1. Các cụm từ này liên kết với nhau bởi dấu +. Ví dụ: Tăng lương + giảm giờ làm, tăng lương + tăng thưởng.

Sau khi đã nhập các từ khóa cần nghiên cứu, người dùng chọn vào nút Tính tần suất, chọn sheet có chứa các câu trả lời gốc rồi chọn tiếp nút OK. Khi đó Module1 sẽ tính toán tần suất các câu trả lời có chứa từng từ khóa trên trong tổng số câu trả lời, đồng thời chuyển các câu trả lời có chứa từ khóa này sang 1 sheet riêng biệt để tiếp tục các phân tích chuyên sâu hơn.

Đối với từ khóa là 1 từ hoặc một cụm từ, tần suất sẽ là các câu trả lời có chứa các từ/ cụm từ này. Đối với các từ khóa là nhóm các cụm từ thì tần suất sẽ là các câu trả lời chứa đồng thời các từ/ cụm từ trong nhóm và các từ/ cụm từ thành phần này không nhất thiết phải đứng cạnh nhau.

**Hình 1:** Tổng hợp các câu trả lời theo từ khóa định sẵn




*Module2: Tìm kiếm các từ khóa mới*

Một trong những kỳ vọng lớn nhất của người nghiên cứu đối với câu hỏi mở là thu thập được các quan điểm độc lập, các câu trả lời mới khác với các phương án định sẵn. Module2 được thiết kế để tìm kiếm các từ khóa mới thông qua tần suất xuất hiện của từng cụm từ trong các câu trả lời thu được.

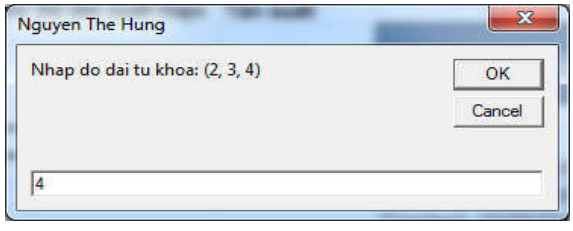
Với giả thiết rằng cụm từ nào xuất hiện nhiều (có tần suất lớn) trong các câu trả lời là các ý kiến phổ biến của người trả lời, Module2 sẽ tính toán tần suất xuất hiện của các cụm từ thông qua độ dài khác nhau của chúng để từ đó xác định được các phương án trả lời mới và phổ biến của người trả lời.

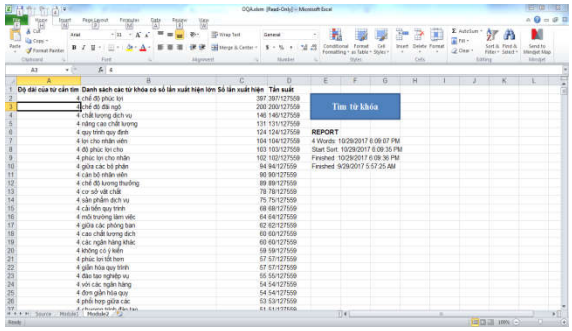
Thông qua các hộp hội thoại thân thiện, Module2 cũng yêu cầu nhập sheet chứa tổng thể câu trả lời cần nghiên cứu, yêu cầu nhập độ dài của từ khóa cần tìm kiếm và số từ khóa có tần suất lớn nhất cần liệt kê. Sau đó, Module2 sẽ chia các câu trả lời thành các cụm từ khác nhau theo độ dài của cụm từ định trước và tính tần suất xuất hiện của từng cụm từ trong toàn bộ các câu trả lời và liệt kê những cụm từ nào có tần suất xuất hiện lớn nhất. Độ dài của các cụm từ này đã loại bỏ các kí tự đặc biệt như: `~!@#$%^&*()-_=+\|]}[{'";:/?.>,< để giảm nhiễu thông tin trong quá trình tổng hợp.

Thông qua các cụm từ có tần suất lớn này, người dùng xác định được các phương án trả lời phổ biến cho câu hỏi nghiên cứu. Đây là những gợi ý quan trọng để người dùng xác định những từ khóa mới, những quan điểm hoặc câu trả lời mới, từ đó sử dụng Module 1 để có các phân tích chuyên sâu hơn.

Ví dụ, đối với câu hỏi xin ý kiến đóng góp của nhiều nhân viên để ngân hàng X phát triển hơn trong năm tới, người dùng khảo sát các cụm từ có độ dài bằng 4 và được kết quả như sau:

**Hình 2:** Tìm kiếm từ khóa thông qua độ dài của cụm từ



Qua kết quả trên, các cụm từ như: "chế độ phúc lợi", "chế độ đãi ngộ", "chất lượng dịch vụ"… là các từ được đề cập đến nhiều nhất. Để nghiên cứu sâu hơn từng cụm từ này, người dùng sử dụng Module1 để thực hiện các công việc tiếp theo.

*Tóm lại,* công cụ OQA chạy trên nền Microsoft Excel nên thân thiện đối với tất cả mọi đối tượng sử dụng khác nhau. Tuy nhiên, công cụ này còn tồn tại và hạn chế là chưa phân tích được ngữ nghĩa của cụm từ trong từng văn cảnh khác nhau. Điều này có thể khắc phục được khi có hệ thống server đủ mạnh để tích hợp công nghệ NLP (Natural language processing) vào công cụ này.

Công cụ OQA cũng có thể xử lý tốt đối với mọi ngôn ngữ (tiếng Anh, tiếng Việt, tiếng Nhật, tiếng Thái…) theo nhiều định dạng phông chữ khác nhau, do vậy có thể áp dụng cho nhiều nghiên cứu trong nước và quốc tế.

Khả năng áp dụng công cụ OQA khá rộng trong việc khai phá dữ liệu dạng chữ trong nhiều lĩnh vực nghiên cứu khác nhau. Những nghiên cứu này không chỉ trong các nghiên cứu về xã hội, nghiên cứu thị trường, khảo sát ý kiến đánh giá... mà còn áp dụng trong các nghiên cứu khác như phân tích định tính để xác định các yếu tố ảnh hưởng đến giá bất động sản thông qua dữ liệu lớn (big data) từ các tin rao quảng cáo trên internet… ***(Xem tiếp trang 28)***

***Tiếp theo trang 31***

**Tài liệu tham khảo:**

1. Bill Jelen and Tracy Syrstad (2010), *VBA and Marcos: Microsoft 2010,* Que Publishing.

2. David Boctor (1999), *Microsoft Office 2000/Visual Basic for Application/ Fundamentals*, Microsoft Office;

3. Robert L. McDonald (2000), *An introduction to VBA in Excel, Finance Dept, Kellogg School,* Northwestern University;

4. Vũ Thị Thu Thủy, Nguyễn Thế Hưng (2016), Ứng dụng ngôn ngữ lập trình VBA trong Excel để giải một số bài toán trong thống kê giá, *Hội thảo khoa học Quốc gia "Thống kê và tin học ứng dụng"*.