Kỹ thuật lập bản đồ và diễn họa các chỉ tiêu thống kê

Haitham Zeidan, Cơ quan thống kê trung ương Palestin (PCBS)

Tóm tắt

Ngày nay, số liệu thống kê ngày càng nhận được nhiều sự quan tâm từ các nhà lập pháp, nhà hoạch định chính sách thành phố, các nhà nghiên cứu và ngay cả người dân. Việc minh họa các dữ liệu dưới dạng dễ hiểu có thể giúp những người ra quyết định hiểu rõ và phân tích một cách hiệu quả lượng thông tin lớn chỉ trong khoảng thời gian ngắn. Diễn họa dữ liệu, một cách diễn đạt nhiều kiểu dữ liệu khác nhau trong một dạng thức dễ hiểu hơn, được áp dụng ngày càng nhiều trong các lĩnh vực. Ngày càng có nhiều cách diễn họa các dữ liệu thống kê, đáng tiếc là mặc dù có nhiều cách diễn họa các dữ liệu thống kê nhưng việc áp dụng chúng vẫn còn hiếm. Hơn nữa, ngay chính việc áp dụng cũng có nhiều hạn chế. Bài viết này nhằm chỉ ra cách thức diễn họa các dữ liệu nhằm cải thiện khả năng đọc và sử dụng dữ liệu thống kê. Các kỹ thuật diễn họa và kỹ thuật tương tác khác nhau đều được sử dụng dựa vào thư viện biểu đồ Highcharts Java Script [1]. Nhằm phân loại nhu cầu người dùng, việc phỏng vấn trước hết sẽ được thực hiện với những người dùng có chuyên môn, cuộc phỏng vấn thứ 2 sau đó được thực hiện để đánh giá việc các kết quả áp dụng cuối cùng kỹ thuật diễn họa của chúng tôi. Kết quả đánh giá đã chỉ ra rằng việc ứng dụng các kỹ thuật diễn họa của chúng tôi đã đạt hiệu quả và hiệu suất cao. Hơn nữa, phần lớn các ứng viên tham gia đánh giá tỏ ra thỏa mãn với việc áp dụng bởi nó giúp họ thực hiện các công việc của họ một cách thành công.

Từ khóa:

Diễn họa dữ liệu, các chỉ số, bản thể học, PCBS, Web 2.0, hợp nhất dữ liệu, kỹ thuật tương tác.

1. Giới thiệu

Các số liệu thống kê chính thức là các số liệu được đưa ra bởi các cơ quan Chính phủ hoặc các tổ chức công khác như các tổ chức quốc tế. Số liệu thống kê cung cấp những thông tin về mặt định lượng lẫn định tính trên tất cả các lĩnh vực chính trong đời sống như kinh tế, phát triển xã hội, các điều kiện sống, y tế, giáo dục và môi trường. Số liệu thống kê chính thức có thể tìm thấy trên các trang web của các cơ quan thống kê quốc gia như: Cơ quan Thống kê trung ương Palestine (PCBS) [2]. Việc có một lược đồ chung để hợp nhất một lượng dữ liệu lớn và diễn họa các kết quả tìm được rất cần thiết để nhờ đó người xem có thể dễ dàng hiểu thông suốt được các dữ liệu. Mục tiêu của việc hợp nhất là làm hài hòa/cân đối các dữ liệu có được từ các nguồn và tài liệu khác nhau, các thuật toán kết nối có thể được sử dụng để kết nối các chỉ số [3].

Triết lý chính của Web 2.0 là hợp tác và chia sẻ, thuật ngữ "Web 2.0" đã tất yếu gắn liền với các dạng phát triển của nó như blog, từ điển mở Wikis, mạng xã hội và sự phát triển của các phần mềm hợp

tác. Web 2.0 đã có những tác động mạnh mẽ trong sự phát triển của những công cụ hợp tác và hợp nhất dùng để phân tích hình ảnh đối với mạng Internet. Các công cụ đó rất cần thiết để tăng cường khả năng của con người trong việc trao đổi những kiến thức thu được đồng thời có thể phát triển những hiểu biết chung với người khác [4].

2. Mục tiêu

Do tính chất phức tạp và không đồng nhất của các số liệu thống kê, nhu cầu có một công nghệ diễn họa hoàn thiện ngày càng cao, chúng tôi đã đưa ra một ứng dụng diễn họa mới nhằm tăng cường khả năng diễn đạt của các số liệu thống kê chính thức dựa trên phương thức phân tích hình ảnh trong đó kết hợp cả việc phân tích dữ liệu và diễn họa tương tác. Chúng tôi đã thực hiện trên các công trình nghiên cứu trước đây về thuật toán kết nối mới [3] dựa trên khoảng cách Hamming [5], khoảng cách diễn giải [6] và bản thể học, bằng cách sử dụng thuật toán của chúng tôi, chúng tôi đã tăng cường sự thống nhất, sự cân bằng và sự kết nối của các chỉ tiêu thống kê từ nhiều nguồn khác nhau, các chỉ số sau khi được nhập đã được lưu lại dưới dạng Khung mô tả tài nguyên (Resource Description Framework - RDF) trong lược đồ, điều này giúp hình ảnh hóa các dữ liệu thống kê bằng cách sử dụng các kỹ thuật diễn họa, lược đồ được xây dựng bao gồm các bảng bản thể học để cải thiện và tăng tính chính xác của thuật toán kết nối. Chúng tôi đã kiểm nghiệm độ chính xác của thuật toán và các kết quả thí nghiệm đã thể hiện sự chính xác cao trong việc kết nối các thuật toán bằng cách bổ sung thêm bản thể học vào thuật toán. Trong bài nghiên cứu này, chúng tôi đã mở rộng công trình của mình để tăng cường sự diễn đạt các số liệu thống kê bằng phương pháp diễn họa và các kỹ thuật tương tác giúp con người có thể phân tích các dữ liệu bằng cách diễn đạt các kết quả dưới dạng trực quan và dễ hiểu khi cho phép tương tác với các dữ liệu. Việc diễn họa các dữ liệu thống kê đã đẩy mạnh việc sử dụng các dữ liệu thống kê giúp hoạt động lập kế hoạch và đưa ra các chính sách được cải thiện. Tất cả các diễn họa đã được thực hiện với sự giúp đỡ từ thư viện Highcharts Java Script [1].

3. Triển khai ứng dụng

Hoạt động diễn họa trong ứng dụng của chúng tôi đã được thực hiện trên nền tảng Microsoft và .NET Framework, sử dụng các công cụ phát triển phần mềm của Microsoft visual studio, bao gồm cả .NET và ASP.NET., thư viện Highcharts Java Script [1], Highcharts là một thư viện biểu đồ được viết hoàn toàn trên nền JavaScript, đã đưa ra một cách thức dễ dàng để đưa các biểu đồ tương tác vào trang web hoặc các ứng dụng web. Highcharts hiện hỗ trợ các dạng đồ thị: đường thẳng, đường cong, lược đồ vùng, lược đồ vùng dạng cong, dạng cột, dạng thanh, hình tròn, dạng điểm, dạng đồng hồ đo, dạng dải vùng, dải vùng cong, dải cột, và dạng cực. Các kỹ thuật tìm kiếm, so sánh, tái dựng, phân loại và chọn lọc tương tác đã được sử dụng để diễn họa trong ứng dụng của chúng tôi. Mục đích của ứng dụng của chúng tôi để lưu trữ dữ liệu và các chỉ số để nhằm sắp xếp, lưu trữ và diễn đạt các dữ liệu một cách đồng nhất và diễn đạt thông tin dưới dạng bảng, biểu và bản đồ, và để tạo điều kiện cho việc chia sẻ dữ liệu được dễ dàng. Ứng dụng này cũng hỗ trợ một số công việc trong việc diễn họa.

4. Đánh giá

Phần này miêu tả các tiếp cận đã được sử dụng để chỉ ra những yêu cầu của người dùng và để đánh giá việc ứng dụng diễn họa. Bước đầu tiên là

thực hiện phỏng vấn để ghi nhận các yêu cầu của người dùng trước khi triển khai các giai đoạn của kỹ thuật diễn họa cho ứng dụng của chúng tôi từ người người dùng cuối cùng. Bước thứ hai là thực hiện phỏng vấn để tiếp nhận và tìm ra những yêu cầu và phản hồi của người dùng cuối cùng đối với việc ứng dụng và các kỹ thuật diễn họa đã được sử dung. Các phần tiếp theo sẽ miêu tả quá trình đánh giá một cách chi tiết hơn.

4.1. Phỏng vấn chuyên gia để đưa ra các yêu cầu chung của người dùng

Tại bước này, ta sẽ thực hiện phỏng vấn để tiếp nhận những yêu cầu trước khi triển khai các giai đoạn của quá trình áp dụng diễn họa từ người dùng cuối cùng, chúng tôi đã thu thập phản hồi bằng việc phỏng vấn 12 ứng viên. Các ứng viên đã được lựa chọn đại diện cho nhóm người dùng có mục đích, bao gồm: các nhà thống kê, nhà nghiên cứu, những người ra quyết định nhờ đó họ có thể đưa ra một cách cụ thể những yêu cầu của người dùng bao gồm cả diễn họa, thiết kế các chứng năng thích hợp, xây dựng hệ thống diễn họa theo yêu cầu của người dùng. Kết quả chúng tôi rút ra từ cuộc phỏng vấn này được tổng kết như sau:

- Tất cả người dùng nói rằng họ quan tâm tới việc diễn họa, hầu hết người dùng quan tâm tới biểu đồ dạng đường thẳng, dạng hình tròn, dạng thanh, dạng cột và dạng bản đồ. 9/12 người dùng đề cập rằng biểu đồ đường thẳng là cách tiện lợi hơn cả để tìm kiếm kết quả.

- Cả chuyên gia và người dùng cuối cùng đều quan tâm tới việc diễn họa các dữ liệu thống kê và họ nghĩ nó có thể giúp công việc của họ dễ dàng hơn và có thể giúp họ hiểu các kết quả rõ ràng và tốt hơn.

- Họ đề nghị sử dụng nhiều dạng hình và biểu đồ khác nhau để miêu tả số liệu thống kê và so sánh các chuỗi thời gian; hệ thống diễn họa nên hỗ trợ nhiều ngôn ngữ khác nhau; sử dụng bản đồ và các màu; bổ sung thêm các hiệu ứng để tải về và chia sẻ các kết quả diễn họa và họ cũng đề xuất sử dụng hiệu ứng hoạt họa cho các chuỗi thời gian và sử dụng các dải màu; linh hoạt hơn trong việc phân loại theo dữ liệu chuỗi thời gian và xem xét các mục và tiểu mục.

- Có 11 người dùng quan tâm đến đối chiếu các kịch bản khác nhau của dữ liệu thống kê, đây có thể là một đặc trưng quan trọng của mô hình diễn họa. Họ nghĩ những đặc trưng này có thể giúp họ đánh giá các kết quả đầu ra và giúp quá trình đưa ra quyết định thuận lợi hơn.

Tất cả người dùng nói rằng họ quan tâm tới việc diễn họa, hầu hết người dùng quan tâm tới biểu đồ dạng đường thẳng, dạng hình tròn, dạng thanh, dạng cột và dạng bản đồ. 9/12 người dùng đề cập rằng biểu đồ đường thẳng là cách tiện lợi hơn cả để tìm kiếm kết quả.

Cả chuyên gia và người dùng cuối cùng đều quan tâm tới việc diễn họa các dữ liệu thống kê và họ nghĩ nó có thể giúp công việc của họ dễ dàng hơn và có thể giúp họ hiểu các kết quả rõ ràng và tốt hơn.

- Họ đề nghị sử dụng nhiều dạng hình và biểu đồ khác nhau để miêu tả số liệu thống kê và so sánh các chuỗi thời gian; hệ thống diễn họa nên hỗ trợ nhiều ngôn ngữ khác nhau; sử dụng bản đồ và các màu; bổ sung thêm các hiệu ứng để tải về và chia sẻ

các kết quả diễn họa và họ cũng đề xuất sử dụng hiệu ứng hoạt họa cho các chuỗi thời gian và sử dụng các dải màu; linh hoạt hơn trong việc phân loại theo dữ liệu chuỗi thời gian và xem xét các mục và tiểu mục.

Có 11 người dùng quan tâm đến đối chiếu các kịch bản khác nhau của dữ liệu thống kê, đây có thể là một đặc trưng quan trọng của mô hình diễn họa. Họ nghĩ những đặc trưng này có thể giúp họ đánh giá các kết quả đầu ra và giúp quá trình đưa ra quyết định thuận lợi hơn.

4.2. Phỏng vấn người dùng cuối để đánh giá kỹ thuật tương tác và diễn họa trong ứng dụng

Một cuộc phỏng vấn với những câu hỏi được đặt ra từ trước đã được thực hiện. 12 ứng viên đã được yêu cầu sử dụng các kỹ thuật diễn họa của ứng dụng để thực hiện một số công việc được lựa chọn từ trước liên quan đến công việc của các ứng viên. Các ứng viên đã được yêu cầu nói ra các suy nghĩ của mình khi thực hiện các công việc. Cuối cùng, các ứng viên được yêu cầu điền vào bảng câu hỏi về toàn bộ việc sử dụng ứng dụng và ước lượng câu trả lời của mình từ mức 1- Hoàn toàn không đồng ý tới mức 5 - Hoàn toàn đồng ý. Tất cả các câu hỏi đều được lựa chọn kỹ lưỡng nhằm đảm bảo đưa ra kết luận về hiệu quả và hiệu suất của ứng dụng, và mức độ hài lòng của người dùng. Tất cả người dùng đó là nhà thống kê, nhà nghiên cứu, những người ra quyết định từ đó họ có thể thực sự ước đoán ứng dụng diễn họa mới này có hiệu quả đủ để giải quyết những công việc nhất định trong lĩnh vực nghiên cứu đó. Trong bảng câu hỏi phản hồi, cũng có một số câu hỏi mở trong đó người tham gia đánh giá có thể viết ra các ý kiến bình luận của của bản thân. Câu trả lời của các câu hỏi này rất quan trọng.

Với kết quả từ bảng câu hỏi phản hồi việc tổng kết các kết quả của cuộc phỏng vấn được kết thúc, những kết luận sau đây về hiệu năng, hiệu quả và sự hài lòng của người dùng đối với ứng dụng:

- Gần 83% người tham gia đều tuyệt đối đồng ý rằng việc sử dụng ứng dụng giải quyết dễ dàng các công việc. Đồng thời những người này cũng chỉ ra rằng họ cảm thấy tự tin về các kết quả họ nhận được sau khi một số công việc hoàn thành.

- Gần 67% người tham gia tuyệt đối đồng ý và thích sự tổng hợp các mặt khác nhau trong cấu trúc của dữ liệu.

- Gần 82% người tham gia không đồng ý rằng họ dành nhiều thời gian để hoàn thành các công việc hoặc rằng họ thường bị nhầm lẫn trong quá trình hoàn thành công việc.

Đồng thời có một số yêu cầu về các tính năng bổ sung:

- Sẽ tốt nếu ứng dụng có thể thực hiện một số phân tích thống kê cơ bản.

- Sẽ rất hữu ích nếu có thêm lựa chọn để bổ sung thêm hơn 1 chỉ số trong cùng một biểu đồ (nếu có thể ứng dụng được và có thể được thực hiện), như khi chúng ta cần chỉ ra những giá trị trong chuỗi thời gian cho số hộ gia đình và quy mô trung bình hộ gia đình trong cùng 1 năm, hoặc dân số và số người thất nghiệp).

- Hỗ trợ nhiều ngôn ngữ khác nhau

- Hiệu ứng hoạt họa và dải màu.

- Thêm nhiều hướng dẫn sử dụng công cụ để giải thích nghĩa của các chỉ số.

5. Kết luận và công trình nghiên cứu trong tương lai

Nghiên cứu này nhằm giới thiệu một ứng dụng diễn họa mới để diễn đạt dữ liệu thống kê. Chúng tôi đã tăng cường việc diễn đạt dữ liệu thống kê trên cơ sở giao diện hình ảnh người dùng năng động và các nguyên lý của phép phân tích hình ảnh (Visual Analytics). Ứng dụng này được giới thiệu nhằm cung cấp những kỹ thuật giúp con người có thể phân tích dữ liệu bằng cách diễn đạt các kết quả một cách trực quan và dễ hiểu trong khi vẫn cho phép có sự tương tác giữa các dữ liệu. Các dữ liệu thống kê được diễn họ giúp đẩy mạnh việc sử dụng các số liệu thống kê trong cải thiện việc lập kế hoạch và xây dựng chính sách. Tất cả các diễn họa được thực hiện với sự hỗ trợ từ thư viện Highcharts Java Script, nhờ đó giúp tạo ra sư tương tác cao hơn và do đó phản hồi tốt hơn cho việc diễn họa của người dùng cuối cùng.

Ứng dụng đã được đánh giá thành công bởi những người dùng khác nhau và các chuyên gia, nhà thống kê, nhà nghiên cứu, những người ra quyết định. Các kết quả đánh giá đã thể hiện tính hiệu quả, hiệu năng ở mức độ cao và sự hài lòng của người dùng. Công trình trong thời gian tới cải thiện sự cộng tác của ứng dụng. Những phương pháp bổ sung sẽ được yêu cầu để hỗ trợ người dùng trong việc tìm kiếm những cách nhìn tốt về dữ liệu và trong việc xác định kỹ thuật diễn họa phù hợp. Chúng tôi sẽ xem xét việc diễn họa 3D đối với các cấu trúc đồ thị không xác định với những thuộc tính không xác định mà chúng tôi nghĩ sẽ vẫn còn là một thử thách ghê gớm.

Tài liệu tham khảo:

[1] Highcharts library written in pure JavaScript: http://www.highcharts.com.

[2] Palestinian Central Bureau of Statistics (PCBS): http://www.pcbs.gov.ps.

[3] H. Zeidan, R. Jayousi and J. Najjar: Interoperable Visualization Framework Towards Enhancing Mapping and Integration of Official Statistics, European Conference on Quality in Official Statistics (Q2014), 2014.

[4] J. Thomas and K. Cook: Illuminating the Path: Research and Development Agenda for Visual Analytics, 2005.

[5] Wikipedia, the free encyclopedia: Hamming Distance: (http://en.wikipedia.org/wiki/Hamming_distance).

[6] Wikipedia, the free encyclopedia: Edit Distance: (http://en.wikipedia.org/wiki/Edit_distance).