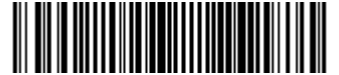




(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN
GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ**

(11)



2-0002536

(51)⁷ **H04L 63/123**

(13) **Y**

(21) 2-2016-00318

(22) 14/09/2016

(45) 25/12/2020 393

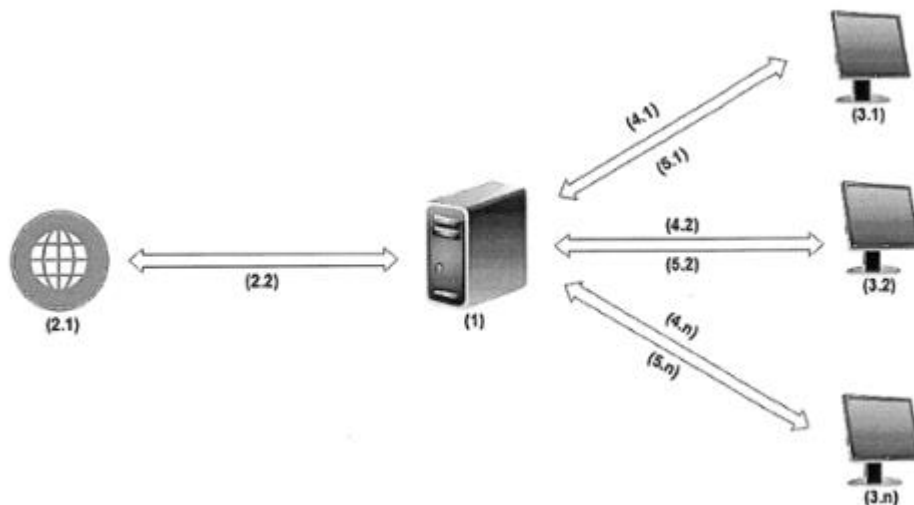
(43) 26/03/2018 360A

(73) Trường Đại học Bách Khoa - Đại học Quốc gia Thành Phố Hồ Chí Minh (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

(72) Đặng Trần Khánh (VN); Nguyễn Thanh Tùng (VN); Trương Quang Hải (VN); Lê Thị Bảo Thu (VN).

(54) PHƯƠNG PHÁP KIỂM TRA DỮ LIỆU DO NHIỀU MÁY CHỦ THU THẬP ĐƯỢC CÓ BỊ THAY ĐỔI DỰA VÀO ĐIỀU PHỐI VÀ SO TRÙNG DỮ LIỆU KHI TIẾN HÀNH THU THẬP TRÊN CÁC TRANG WEB THƯƠNG MẠI ĐIỆN TỬ

(57) Giải pháp hữu ích đề cập đến phương pháp kiểm tra dữ liệu do nhiều máy chủ thu thập được có bị thay đổi hay không dựa vào điều phối và so trùng dữ liệu rút trích nhằm kiểm tra tính chính xác cho quá trình thu thập dữ liệu trên các trang web thương mại điện tử. Với việc có một máy chủ master (máy chủ điều phối) đóng vai trò điều phối và quản lý việc rút trích dữ liệu ở các máy chủ slave (nhân công), ta có thể tiến hành rút trích cùng một lúc nhiều trang chi tiết sản phẩm và phát hiện được những thông tin quan trọng đã bị thay đổi. Song song đó, bằng chính sách phân chia thông tin sản phẩm cần rút trích một cách hợp lý, số lượng dữ liệu thu thập giống nhau ít nhất có thể, đảm bảo được hiệu suất của quá trình này. Nhờ vậy, nguồn dữ liệu thu thập được chính xác nhất có thể, phục vụ các mục đích phân tích về sau.



Lĩnh vực kỹ thuật được đề cập

Giải pháp hữu ích thuộc lĩnh vực khoa học máy tính, chuyên ngành hệ thống thông tin, ứng dụng vào việc kiểm tra dữ liệu rút trích được từ các trang web thương mại điện tử xem dữ liệu có bị thay đổi hay không, bằng cách điều phối việc rút trích dữ liệu giữa nhiều nhà cung cấp dịch vụ một cách hợp lý và tiến hành so trùng kết quả trả về. Chúng ta sẽ chỉ cho phép một vài tài nguyên, thông tin quan trọng của trang web được cùng lúc nhiều máy chủ thu thập. Do đó, có thể giảm chi phí rút trích xuống tối đa, mà vẫn đảm bảo những thông tin chính yếu được chính xác. Mục đích của lĩnh vực kỹ thuật được đề cập là nhằm hạn chế nhiều nhất có thể những dữ liệu rút trích được từ các trang web thương mại điện tử đã bị thay đổi làm sai lệch, do nhà cung cấp dịch vụ gian lận hoặc do chính trang web thương mại điện tử không muốn người khác thu thập dữ liệu đúng của mình.

Tình trạng kỹ thuật của giải pháp hữu ích

Trong mô hình xác thực dữ liệu thu thập truyền thống, cách đơn giản nhất để doanh nghiệp biết được một nhà cung cấp dịch vụ thu thập dữ liệu có gian lận, sửa đổi dữ liệu trước khi cung cấp dữ liệu cho mình hay không, chính là thuê một nhà cung cấp dịch vụ thu thập dữ liệu khác. Họ sẽ thu thập dữ liệu từ các trang web giống như nhà cung cấp dịch vụ thu thập dữ liệu ban đầu. Chủ doanh nghiệp lúc này chỉ cần so sánh dữ liệu từ hai nhà cung cấp dịch vụ thu

thập dữ liệu trả về. Nếu có sự khác nhau giữa hai nhà cung cấp dịch vụ này thì nguy cơ dữ liệu đã được thay đổi bởi một trong hai nhà cung cấp dịch vụ thu thập dữ liệu được thuê là rất lớn. Từ đó, niềm tin vào dữ liệu từ hai nhà cung cấp dịch vụ này thu thập được sẽ bị giảm đáng kể. Lúc này, doanh nghiệp cần có các hành động tiếp theo để quyết định có nên tiếp tục sử dụng hai nhà cung cấp dịch vụ này hay không cũng như quyết định các hành động liên quan đến dữ liệu đã thu thập được.

Tuy nhiên, đối với mô hình này, nhược điểm lớn nhất là chủ doanh nghiệp phải bỏ ra gấp đôi chi phí để kiểm tra dữ liệu của mình có bị sửa đổi hay không. Từ đó đặt ra vấn đề là làm sao để có thể phát hiện ra dữ liệu có bị sửa đổi hay không với một chi phí phải bỏ ra là thấp nhất.

Do đó, cần có một giải pháp để lượng dữ liệu rút trích trùng nhau là ít nhất có thể mà vẫn có thể phát hiện ra rằng dữ liệu có bị thay đổi hay không. Để giải quyết vấn đề này, nếu chủ doanh nghiệp biết được dữ liệu nhà cung cấp dịch vụ thu thập dữ liệu thường sửa đổi, gian lận ở những vị trí nào nhiều nhất thì chủ doanh nghiệp chỉ cần rút trích thêm đúng những dữ liệu được xem là quan trọng và có nguy cơ bị sửa đổi. Khi đó, lượng dữ liệu cần phải đi thu thập lại sẽ giảm đáng kể và có thể tiết kiệm rất nhiều chi phí cho việc kiểm tra dữ liệu có bị gian lận hay không.

Ở Việt Nam hiện chưa có bất kỳ kết quả nghiên cứu nào được đăng ký bảo hộ tương tự.

Bản chất kỹ thuật của giải pháp hữu ích

Mục đích của giải pháp hữu ích là nhằm phát hiện những dữ liệu rút trích đã bị thay đổi làm sai lệch từ phía nhà cung cấp dịch vụ rút trích được

thuê. Các đối thủ cạnh tranh (người thuê các nhà cung cấp dịch vụ thu thập dữ liệu sửa đổi dữ liệu của đối thủ - gọi là người tấn công) thường không sửa đổi dữ liệu một cách tùy tiện mà chỉ muốn thay đổi những dữ liệu mà đối với họ được xem là quan trọng, có ảnh hưởng trực tiếp đối với đối thủ của họ và lĩnh vực họ đang hoạt động. Do đó, việc đặt các dữ liệu quan trọng ở mức độ ưu tiên cao là hoàn toàn hợp lý. Tuy nhiên, việc chỉ ra đâu là dữ liệu quan trọng và liệu chỉ những dữ liệu được xem là quan trọng đối với chủ doanh nghiệp đã đủ để xác thực một nhà cung cấp dịch vụ thu thập dữ liệu hay chưa. Có một thực tế tồn tại là đối với một doanh nghiệp thì không những có một đối thủ mà có rất nhiều đối thủ cạnh tranh, mỗi đối thủ lại đóng vai trò như một người tấn công vào dữ liệu. Do đó, việc xác định dữ liệu quan trọng có thể dẫn đến toàn bộ dữ liệu mà chủ doanh nghiệp thuê nhà cung cấp dịch vụ thu thập dữ liệu đều là quan trọng. Và kết quả là sẽ dẫn đến trường hợp như mô hình truyền thống.

Vì vậy cần có một biện pháp xác định tầm quan trọng của dữ liệu và có chính sách rút trích hợp lý để có thể kiểm tra tính chính xác của dữ liệu mà vẫn đảm bảo được chi phí rút trích dữ liệu ở mức thấp nhất.

Từ đó, mục đích của giải pháp hữu ích là nhằm thiết kế một mô hình rút trích và kiểm tra dữ liệu có bị thay đổi hay không mà vẫn đảm bảo được về mặt chi phí, hiệu suất cho doanh nghiệp.

Để đạt được mục đích trên, giải pháp hữu ích được xây dựng thông qua ba bước cơ bản sau:

- Bước 1: Một máy chủ đóng vai trò là master (*máy chủ điều phối*) sẽ rút trích thông tin tất cả các nhóm danh mục sản phẩm. Trong đó, sẽ có

thông tin chung (như định danh, tên, địa chỉ trang web chi tiết sản phẩm,...) của tất cả các sản phẩm mà trang web có.

- Bước 2: Máy chủ master phụ trách phân chia các sản phẩm có được thành từng nhóm. Các nhóm này sẽ có một vài sản phẩm trùng nhau, sao cho số lượng trùng nhau này là ít nhất có thể. Mỗi nhóm sẽ được gửi cho một máy chủ slave (*nhân công*) để rút trích thông tin chi tiết của sản phẩm và trả kết quả về cho máy chủ master.

- Bước 3: Kiểm tra, đánh giá kết quả có gì sai khác giữa cùng một sản phẩm do các máy chủ slave khác nhau thu thập dữ liệu.

Nói tóm lại, giải pháp hữu ích này nhằm kiểm tra tính chính xác mà vẫn đảm bảo được chi phí thấp nhất cho quá trình thu thập dữ liệu trên các trang web thương mại điện tử. Nhờ vậy, hạn chế đến mức tối đa sai lệch trong dữ liệu thu thập được, phục vụ cho các mục đích phân tích về sau.

Mô tả vắn tắt các hình vẽ

Hình 1: Mô hình điều phối rút trích dữ liệu.

Hình 2: Bộ phân loại của máy chủ Master.

Hình 3: Bộ so trùng của máy chủ Master.

Mô tả chi tiết giải pháp hữu ích

Hình 1 mô tả mô hình điều phối rút trích dữ liệu. Một máy chủ (1) đóng vai trò master (*máy chủ điều phối*) chịu trách nhiệm rút trích thông tin chung, như định danh, tên, địa chỉ trang web (hay còn gọi là url - uniform resource locator (*định vị tài nguyên thống nhất*)) chi tiết sản phẩm,... của tất cả sản phẩm hiện có trên trang web (2.1), phân bổ các url cần rút trích dữ liệu (4.1, 4.2, 4.n) cho các máy chủ slave (*nhân công*) khác (3.1, 3.2, 3.n). Các máy chủ

slave tiến hành rút trích những url này cùng một lúc để có thông tin chi tiết của sản phẩm và trả kết quả (5.1, 5.2, 5.n) về cho máy chủ master. Kết quả này sẽ được tiến hành so trùng ở những url giống nhau được nhiều máy chủ slave cùng rút trích. Nếu giống nhau thì dữ liệu rút trích được là chính xác, ngược lại, dữ liệu có thể đã bị thay đổi làm cho sai khác.

Hình 2 mô tả bộ phân loại url của máy chủ master. Dữ liệu đầu vào là tất cả url chi tiết sản phẩm của trang web (2.2), bộ phân loại (4) sẽ chia nhỏ thành những nhóm url (4.1, 4.2, 4.n) để đưa cho các máy chủ slave tương ứng.

Hình 3 mô tả bộ so trùng của máy chủ master. Dữ liệu đầu vào là kết quả rút trích được của cùng một url ở các máy chủ slave khác nhau (5.1, 5.2, 5.n). Bộ so trùng (6) sẽ tiến hành kiểm tra và trả về kết quả so trùng là giống nhau (6.1) hoặc khác nhau (6.2). Thông tin rút trích có kết quả so trùng giống nhau (6.1) được lưu trữ vào cơ sở dữ liệu (7).

Cách thức hoạt động của hệ thống:

- Máy chủ master (1) rút trích thông tin các danh mục sản phẩm hiện có của trang web (2.1). Từ đây, máy chủ master sẽ có được thông tin chung của tất cả các sản phẩm, trong đó có các url của trang chi tiết sản phẩm (2.2). Để phân chia nhiệm vụ thu thập cho các máy chủ slave (3.1, 3.2, 3.n), các url này sẽ được chuyển qua một bộ phân loại (4) (Hình 2) để chia thành các nhóm url (4.1, 4.2, 4.n) (số lượng nhóm bằng số lượng máy chủ slave). Giữa các nhóm sẽ có những url giống nhau. Tức là sẽ có những url được nhiều máy chủ slave cùng rút trích. Những url được xem là quan trọng sẽ có xác suất xuất hiện cao hơn ở nhiều hơn một nhóm. Ngược lại, những url ít

quan trọng, chỉ cần đưa cho một máy chủ slave phụ trách rút trích là được.

- Thông thường, những thông tin quan trọng, mới nhất, được người dùng quan tâm nhất sẽ được đặt ở phía đầu mỗi trang web. Do đó, ta sẽ đánh xác suất được rút trích ở nhiều hơn một máy chủ slave cho những sản phẩm xuất hiện ở đầu mỗi trang danh mục là cao nhất và sau đó giảm dần cho những sản phẩm phía sau. Ví dụ: 10% những sản phẩm đầu tiên ở mỗi danh mục sẽ chắc chắn được chia ở hai nhóm. 20% sản phẩm tiếp theo xác suất được xuất hiện ở hai nhóm là 0.8. 20% tiếp theo nữa, xác suất này còn 0.75,... và cứ thế giảm dần. Nhờ đó, lượng chi tiết sản phẩm cần rút trích trùng nhau sẽ giảm xuống tối đa, mà vẫn đảm bảo những thông tin quan trọng luôn được nhiều máy cùng rút trích độc lập và sau đó tiến hành kiểm tra lại.
- Sau khi hoàn thành bước rút trích, các máy chủ slave sẽ trả kết quả dữ liệu (5.1, 5.2, 5.n) thu thập được, chính là thông tin chi tiết của sản phẩm được phân bổ, về cho master. Kết quả này được đưa vào bộ so trùng (6), để kiểm tra dữ liệu có sai khác hay không. Nếu giống nhau, dữ liệu (6.1) được xem là chính xác, không bị thay đổi, chúng ta sẽ tiến hành lưu vào cơ sở dữ liệu (7). Nếu khác nhau, máy chủ master có thể tiến hành yêu cầu một máy chủ slave khác rút trích lại dữ liệu (6.2), lúc này được xem là không chính xác.
- Tuy nhiên, thông tin trên trang web tại những thời điểm khác nhau có thể khác nhau. Ví dụ: thông tin mô tả sản phẩm hay giá cả được

cập nhật, thông tin quảng cáo tại những lần tải (load) một trang web cũng khác nhau. Do đó, kết quả rút trích của các máy chủ slave có thể khác nhau mà không phải có sự cố tình can thiệp làm sai lệch dữ liệu. Vì vậy, chúng ta sẽ chỉ so sánh một vài thông tin chủ chốt tùy vào mục đích thu thập dữ liệu. Ví dụ dữ liệu trên một trang rao vặt:

- Họ tên người dùng đăng kí: hầu như không thay đổi.
- Số điện thoại: một người có thể có nhiều số điện thoại, chỉ cần giống nhau ít nhất một số.
- Tên sản phẩm, giá sản phẩm: không bị thay đổi thường xuyên trong một khoảng thời gian ngắn.
- Thông tin đánh giá sản phẩm, phản hồi của khách hàng: trùng khớp ở một ngưỡng nào đó (ví dụ: >80%).
- ...

Ví dụ thực hiện giải pháp hữu ích

Giải pháp hữu ích này có thể được tích hợp vào giai đoạn thu thập dữ liệu trong hệ thống rút trích và phân tích dữ liệu tự động từ các trang web thương mại điện tử. Ví dụ, chúng ta cần xây dựng một hệ thống có chức năng trả lời cho khách hàng biết nên quảng cáo một loại sản phẩm ở trang thương mại điện tử nào là có lợi nhất. Hệ thống này cần thu thập tất cả dữ liệu trên các trang thương mại điện tử cho phép rao vặt, quảng cáo (như raovat, chợ tốt, lazada,...). Tiến hành phân tích trên những dữ liệu này sẽ có kết quả nên quảng cáo một sản phẩm nào đó trên trang web nào sẽ nhanh bán được giá cao hơn, thu hút được nhiều quan tâm của khách hàng hơn. Tuy nhiên, trong trường hợp các trang web rao vặt trên vì lý do cạnh tranh hoặc không muốn người khác thu

thập dữ liệu của mình hoặc các nhà cung cấp dịch vụ rút trích mà hệ thống thuê đã thay đổi dữ liệu, sẽ làm cho nguồn dữ liệu bị sai lệch. Do đó, kết quả phân tích cũng không còn chính xác nữa.

Tích hợp giải pháp kiểm tra dữ liệu bị thay đổi được trình bày trong giải pháp hữu ích, ta có thể đảm bảo dữ liệu thu thập được không bị thay đổi. Đồng thời, hiệu suất của quá trình này cũng được đảm bảo. Kết quả phân tích sẽ không bị nhiễu bởi sự không chính xác của dữ liệu.

Những lợi ích giải pháp đạt được

Việc thực hiện giải pháp hữu ích này, đem lại một số lợi ích như sau:

- Tránh được việc thu thập dữ liệu bị sai do có sự thay đổi từ nhà cung cấp dịch vụ hoặc từ chính trang web thương mại điện tử. Sử dụng cùng lúc nhiều nhà cung cấp dịch vụ thu thập dữ liệu và tiến hành đối chiếu kết quả có được giúp doanh nghiệp có thể xác định những sai khác và khả năng bị thay đổi dữ liệu. Từ đó, đảm bảo dữ liệu thu thập được là chính xác, không bị can thiệp dẫn đến bị nhiễu, ảnh hưởng đến kết quả phân tích sau này.

- Giảm thiểu chi phí do phải thuê cùng lúc nhiều nhà cung cấp tiến hành thu thập những dữ liệu giống nhau. Lượng dữ liệu giống nhau này là ít nhất có thể. Vì vậy, có thể giảm tối đa chi phí rút trích dữ liệu.

- Ngoài ra, chúng ta sẽ cho phép nhiều máy chủ cùng lúc tiến hành thu thập, mỗi máy thu thập một phần của dữ liệu. Cho nên, thời gian cho công việc này sẽ được rút ngắn lại.

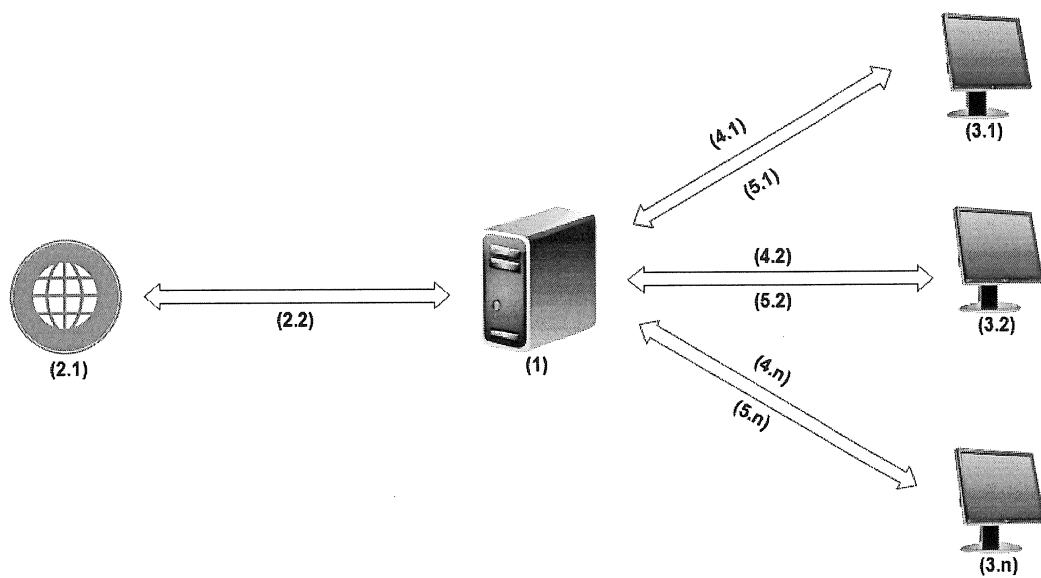
YÊU CẦU BẢO HỘ

1. Phương pháp kiểm tra dữ liệu do nhiều máy chủ thu thập được có bị thay đổi hay không dựa vào điều phối và so trùng dữ liệu rút trích gồm ba bước:

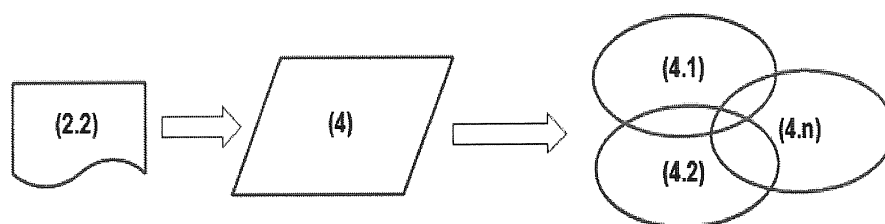
(i) một máy chủ đóng vai trò là master (*máy chủ điều phối*) sẽ rút trích thông tin chung gồm: định danh, tên, địa chỉ trang web (hay còn gọi là url - uniform resource locator (*định vị tài nguyên thống nhất*) chi tiết sản phẩm của tất cả các sản phẩm hiện có của trang web;

(ii) máy chủ master phân chia tất cả các sản phẩm (hay là các url chi tiết sản phẩm) thành từng nhóm bởi bộ phân loại sao cho số lượng sản phẩm giống nhau giữa các nhóm là ít nhất có thể bằng cách đặt xác suất được rút trích ở nhiều hơn một máy chủ slave (nhân công) cho những sản phẩm xuất hiện ở đầu mỗi trang danh mục là cao nhất và sau đó đặt xác suất được rút trích giảm dần cho những sản phẩm phía sau, sau đó mỗi nhóm sẽ được gửi cho một máy chủ slave để rút trích thông tin chi tiết của sản phẩm và trả kết quả về cho máy chủ master;

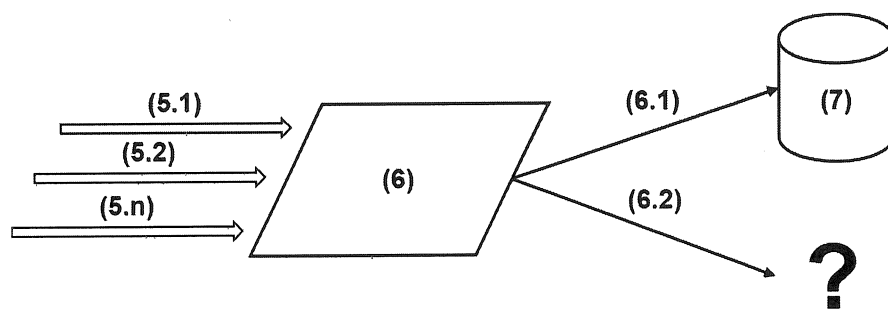
(iii) kiểm tra, đánh giá kết quả có gì sai khác giữa cùng một sản phẩm do các máy chủ slave khác nhau thu thập dữ liệu bởi bộ so trùng dữ liệu của máy chủ master.



Hình 1



Hình 2



Hình 3